

2018 EE448, Big Data Mining, Lecture 12

Real-Time Bidding & Behavioral Targeting

Weinan Zhang

Shanghai Jiao Tong University

<http://wnzhang.net>

<http://wnzhang.net/teaching/ee448/index.html>

Content of This Course

- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

Display Advertising

Exxon Mobil Investigated in New York Over Possible Lies on Climate

By JUSTIN GILLIS and CLIFFORD KRAUSS
3:30 PM ET

The sweeping inquiry, by the state attorney general, focuses on whether the oil company lied to the public and investors over the risks of climate change.

250 Comments



T. Fallon/Bloomberg, via Getty Images

An Exxon Mobil refinery in Los Angeles, Calif. The New York attorney general is investigating the oil and gas company.

LATEST NEWS

- 5:01 PM ET 'Grand Theft Auto' Maker Take-Two's Revenue Nearly Triples
- 5:00 PM ET United Airlines CEO to Return in Early 2016 After Heart Attack
- 4:57 PM ET NY Attorney General Investigating Exxon Over Climate Statements

MARKETS »

At close 11/05/2015

European Union Predicts Economic Gains From Influx of Migrants

By JAMES KANTER
12:10 PM ET

Officials forecast that the three million arrivals expected by 2017 would provide a net gain of perhaps a quarter of 1 percent by that year to the European economy.



INSIGHT & ANALYSIS

COMMON SENSE

Dewey Jury's Deadlock Exposes a System's Flaws

By JAMES B. STEWART
3:06 PM ET

One reason for the mistrial in the Dewey & LeBoeuf criminal case may have been the requirement for a unanimous decision.

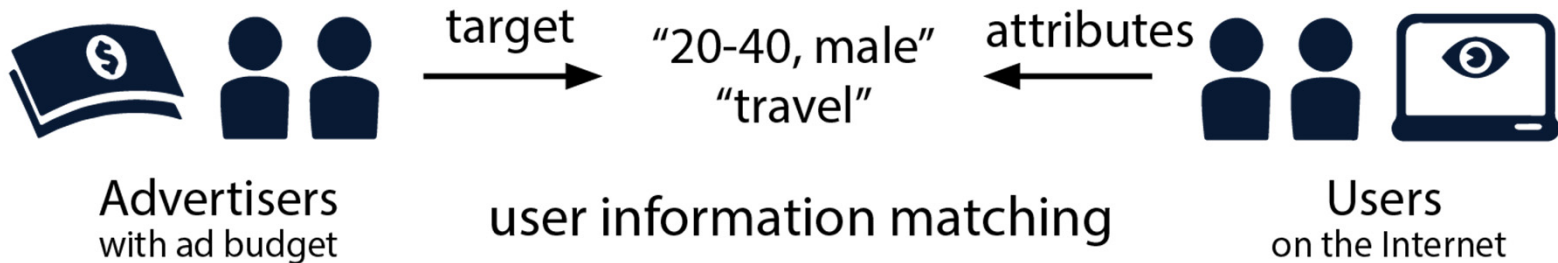


 **BACKBASE**

Backbase a Leader in the Forrester Wave for Omni-Channel Digital Banking

Read the Report

Display Advertising



- Advertiser targets a segment of users
 - No matter what the user is searching or reading
- Intermediary matches users and ads by user information

Internet Advertising Frontier:

Real-Time Bidding (RTB) based Display Advertising

What is Real-Time Bidding?

- Every online **ad view** can be evaluated, bought, and sold, all **individually**, and all **instantaneously**.
- Instead of buying keywords or a bundle of ad views, advertisers are now **buying users** directly.
- **Behavioral targeting**: it is possible now to **track** user actions resulted from an online campaign, advertising optimization becomes more resembling to that of the financial market trading and tends to be driven by the marketing profit and return-on-investment (ROI).

An Example of RTB

Suppose a student regularly reads articles on emarketer.com

The screenshot shows the eMarketer website interface. At the top is a navigation bar with the eMarketer logo and links for Research Topics, Products, Why eMarketer, Customer Stories, and Articles. The main article is titled "Advertisers Continue Rapid Adoption of Programmatic Buying" with a sub-headline "By 2017, advertisers will spend more than \$9 billion on RTB". The article date is Nov 26, 2013, and there are share, print, and email options. The article text discusses the growth of RTB and includes a bar and line chart showing RTB digital display ad spending, percentage change, and percentage of total digital display ad spending from 2012 to 2017. A sidebar on the right lists "Latest from eMarketer" with links to various articles and a "Contact Sign-Up" button. A content-related advertisement is highlighted with a red box, featuring the text "MARKETING PROGRAMS FOR EMAIL MARKETERS" and "FREE DOWNLOAD", with buttons for "WATCH THE VIDEO." and "DO WHAT CAN NOW BE DONE."

Advertisers Continue Rapid Adoption of Programmatic Buying
By 2017, advertisers will spend more than \$9 billion on RTB

Nov 26, 2013

Share Print Email

Advertisers are spending more than expected on real-time bidding, which is expected to account for a significant share of all display ad spending in the US advertising — which includes RTB — continues its rapid transition from infancy to a well-established display purchase method in just a few years.

eMarketer projects RTB digital display ad spending in the US will account for 29.0% of total US digital display ad spending by 2017, or \$9.03 billion. In 2013, it will account for 19.0%, or \$3.37 billion. These estimates are revised slightly upward from our previous forecast in August

Year	RTB digital display ad spending (billions)	% change	% of total digital display ad spending
2012	\$1.92	94.8%	13.0%
2013	\$3.37	75.3%	19.0%
2014	\$4.66	38.4%	22.0%
2015	\$6.15	31.9%	25.0%
2016	\$7.83	28.0%	27.4%
2017	\$9.03	15.3%	29.0%

Note: includes all display formats served to all devices
Source: eMarketer, Dec 2013
166097 www.emarketer.com

MARKETING PROGRAMS FOR EMAIL MARKETERS
FREE DOWNLOAD

WATCH THE VIDEO.
DO WHAT CAN NOW BE DONE. ©

Content-related ads

An Example of RTB

He recently checked the London hotels

Booking.com ? £ US Recently viewed Lists 3 Weinan Zhang B

Browse by destination theme [Shopping](#) [Fine Dining](#) [Culture](#) [Sightseeing](#) [Monuments](#) [Relaxation](#)

[home](#) → [uk](#) → [greater london](#) → [london](#) → search results (In fact, no login is required)
16,378 properties 1,824 properties 1,574 properties London, 2 adults, 11 nights (Jul 14 - Jul 25) [Change dates](#)

Search

Destination/Hotel Name:

Distance:

Check-in Date

Check-out Date

I don't have specific dates yet

Guests

Search
Search properties


London is a top choice with fellow travelers on your selected dates (48% reserved).
Tip: Prices might be higher than normal, so try searching with different dates if possible.




48% reserved

[Try previous week](#) Jul 7 - Jul 18 [Try next week](#) Jul 21 - Aug 1


930 out of 1857 properties are available in and around London
Showing 1 – 15


Sort by: **Recommended** Stars Location Price Review Score [List](#) [Map](#)





Park Plaza Victoria London ★★★★★   1736
[Central London, Westminster, London](#) •  **Nearby stop**

There are 13 people looking at this hotel.
Latest booking: 1 hour ago

 Superior Double Room **We have 5 rooms left!** **£2,353.65** **Price for 11 nights**
7 more room types [Book now](#)



Central Park Hotel ★★★   1993 **6.6**

Filter by:

An Example of RTB

Relevant ads on [facebook.com](https://www.facebook.com)

The image shows a screenshot of a Facebook homepage. The top navigation bar includes the Facebook logo, a search bar, and the user's name 'Weinan' with 'Home' and other navigation icons. The left sidebar contains a list of friends and groups, including 'Family', 'UCL', 'SJTU', and 'Microsoft Research C...'. The main content area features a sponsored advertisement for 'Secret Escapes' with a red border. The ad includes the Secret Escapes logo, the text 'Sponsored · ✨', a 'Like Page' button, and the headline 'Find the best rates on handpicked hotels'. Below the headline is a large image of a modern, illuminated outdoor lounge area with a pool. The ad text reads 'Secret Escapes | Exclusive Discounts' and 'Get up to 70% off luxury hotels and holidays.', with a 'Sign Up' button and the website 'WWW.SECRETESCAPES.COM'. Below the ad are engagement metrics: 'Like · Comment · Share · 2,327 85 444'. To the right of the Secret Escapes ad is a friend's profile for 'Bingkai Lin' and 'Zhaomeng Peng'. Below these are two more sponsored advertisements, also with red borders. The first is for '247 London Hostel' with the text 'SPONSORED', 'See all', '247 London Hostel', 'booking.com', and 'Book & Save! 247 London Hostel, London.' The second is for 'Stale Marketing Stinks' with the text 'SPONSORED', 'emarketer.com', the eMarketer logo, and 'Freshen up with eMarketer's reports, trends & data on digital marketing. Download Today!'. At the bottom right, there is a footer with 'English (UK) · Privacy · Terms · Cookies · More ▾'.

An Example of RTB

Even on supervisor's homepage!
(User targeting dominates the context)

DR. JUN WANG
Computer Science, UCL



About Me

Contact

Publications

Teaching

Research Team

Prospective Students

Type text to search here...



CIKM2013 Tutorial: Real-Time Bidding: A New Frontier of Computational Advertising Research

July 30th, 2013

Comments of

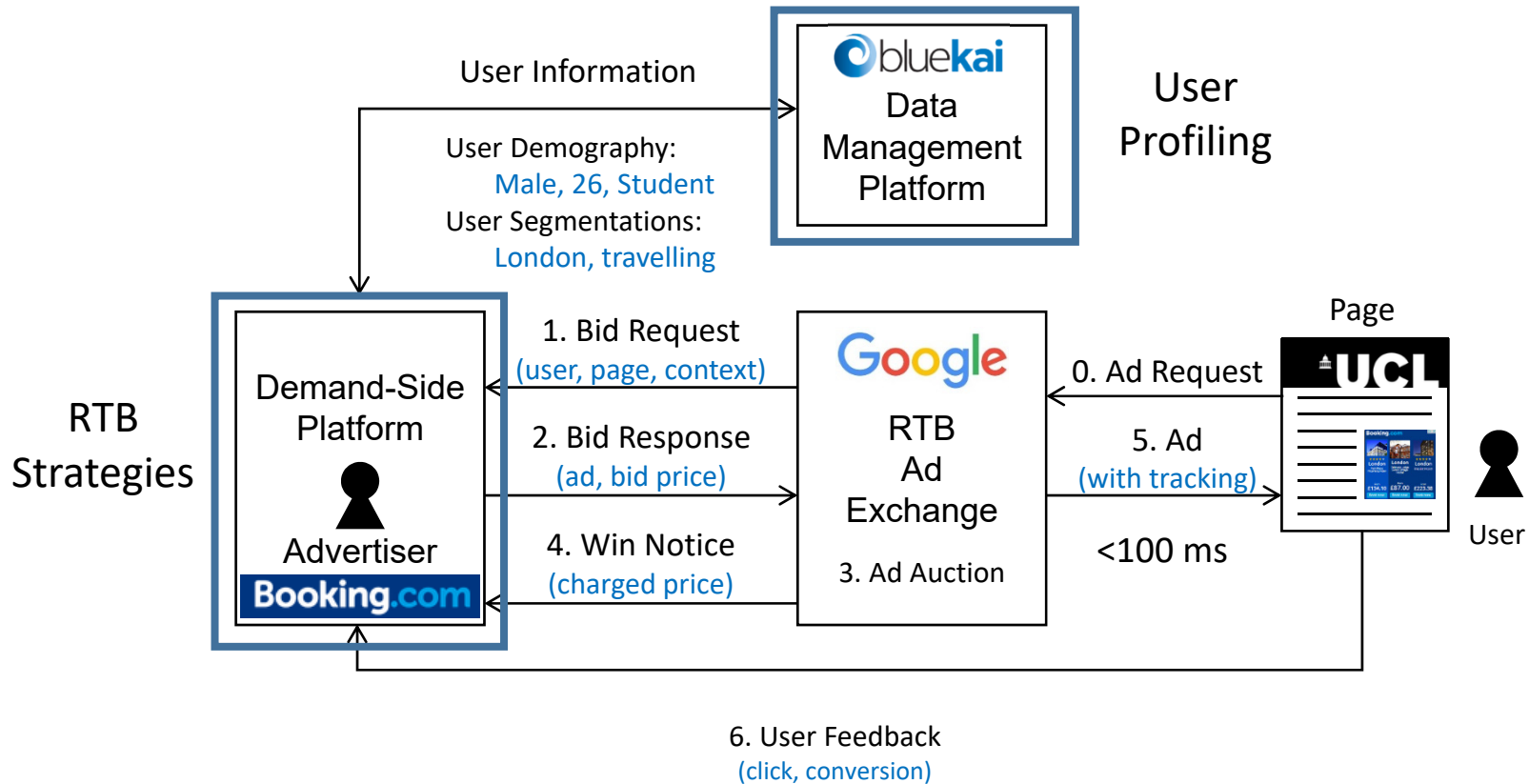
Online advertising is now one of the fastest advancing areas in IT industry. In display and mobile advertising, the most significant development in recent years is the growth of Real-Time Bidding (RTB), which allows selling and buying online display advertising in real-time one ad impression at a time. Since then, RTB has fundamentally changed the landscape of the digital media market by scaling the buying process across a large number of available inventories. It also encourages behaviour (re-)targeting, and makes a significant shift toward buying focused on user data, rather than contextual data. A report from IDC shows that in 2011, global RTB based display ad spend increased by 237% compared to 2010, with the U.S.'s \$2.2 billion RTB display spend leading the way. The market share of RTB-based spending of all display ad spending will grow from 10% in 2011 to 27% in 2016, and its share of all indirect spending will grow from 28% to 78%.

Scientifically, the further demand for automation, integration and optimization in RTB brings new research opportunities in the CIKM fields. For instance, the much enhanced flexibility of

“Relevant” Ads or not?

Hotel Name	Star Rating	Price (From)
Park Plaza Victoria London	5 stars	£134.10
Palmers Lodge Swiss Cottage Hostel	3 stars	£87.00
Thistle Hotel	5 stars	£223.38

RTB Display Advertising Mechanism



- Buying ads via real-time bidding (RTB), 10 billion per day
- A real big data battlefield

RTB: A Big Data Battle Field

- The daily volume of RTB platforms and the comparison with finance institutes

	DSP/Exchange	Daily Traffic
Advertising	iPinYou, China	18 billion impressions
	YOYI, China	5 billion impressions
	Fikisu, US	32 billion impressions
Finance	New York Stock Exchange	12 billion shares
	Shanghai Stock Exchange	14 billion shares

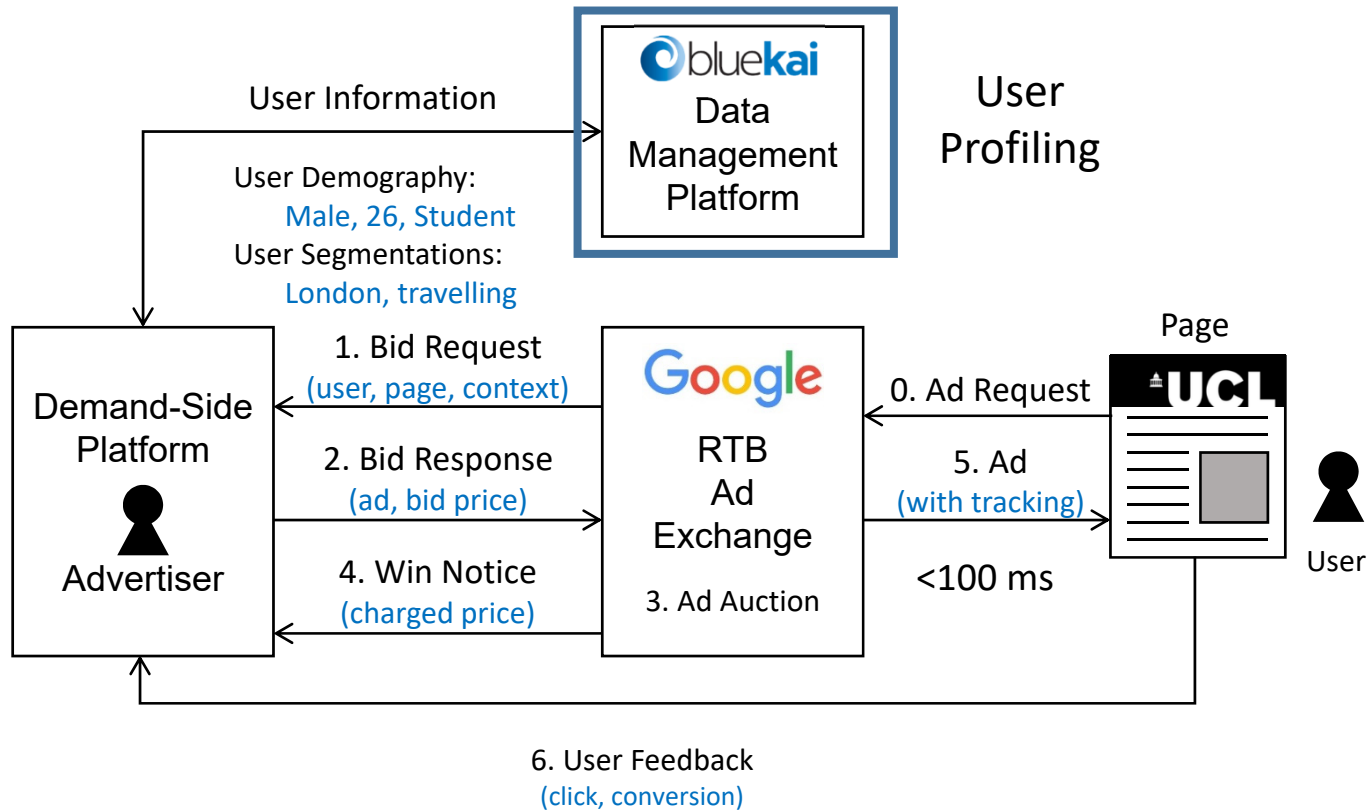
	Query per Second
Turn DSP	1.6 million
Google	40,000 search

It is fair to say that the transaction volume from display advertising has already surpassed that of the financial market

Content of This Course

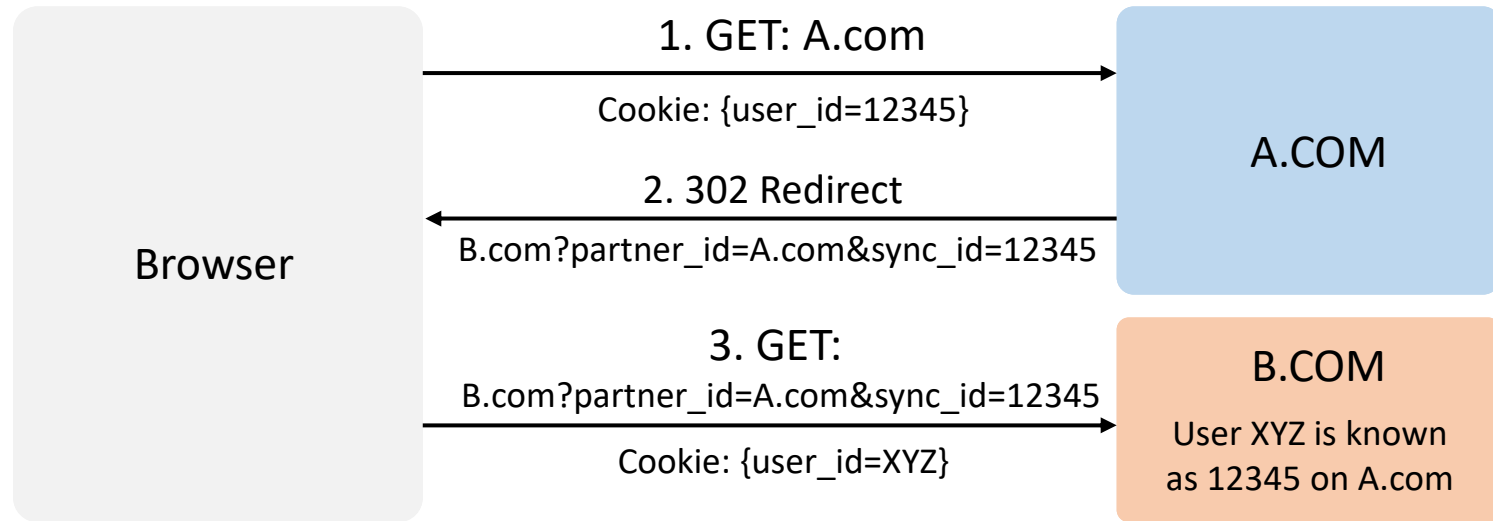
- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

DMP: Data Management Platform



- DMP is a data warehouse that stores, merges, and sorts, and labels it out in a way that's useful for marketers, publishers and other businesses.

Cookie Sync: Merging Audience Data

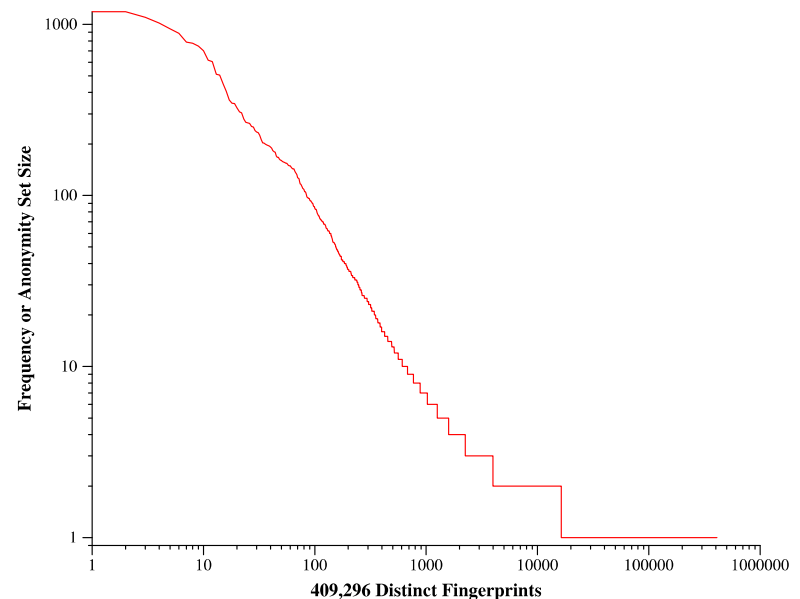


When a user visits a site (e.g. ABC.com) including A.com as a third-party tracker.

- (1) The browser makes a request to A.com, and included in this request is the tracking cookie set by A.com.
- (2) A.com retrieves its tracking ID from the cookie, and redirects the browser to B.com, encoding the tracking ID into the URL.
- (3) The browser then makes a request to B.com, which includes the full URL A.com redirected to as well as B.com's tracking cookie.
- (4) B.com can then link its ID for the user to A.com's ID for the user2

Browser Fingerprinting

- A **device fingerprint** or **browser fingerprint** is information collected about the remote computing device for the purpose of identifying the user.
- Fingerprints can be used to **fully or partially identify individual users** or devices even when cookies are turned off.



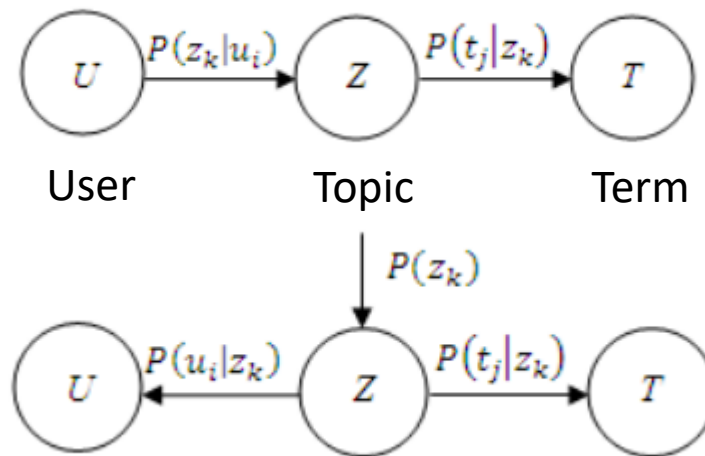
94.2% of browsers with Flash or Java were unique in a study

Eckersley, Peter. "How unique is your web browser?." Privacy Enhancing Technologies. Springer Berlin Heidelberg, 2010.

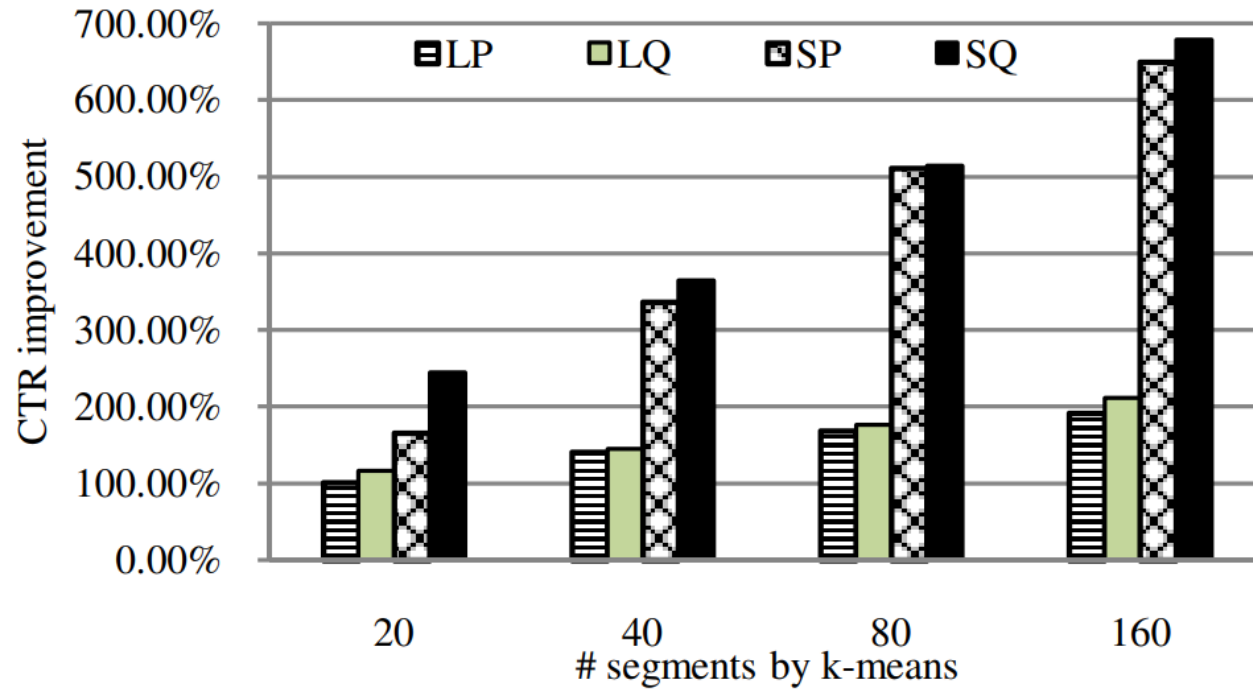
Acar, Gunes, et al. "The web never forgets: Persistent tracking mechanisms in the wild." Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014.

User Segmentation and Behavioral Targeting

- Behavioral targeting helps online advertising
- From user – documents to user – topics
 - Latent Semantic Analysis / Latent Dirichlet Allocation



User Segmentation and Behavioral Targeting

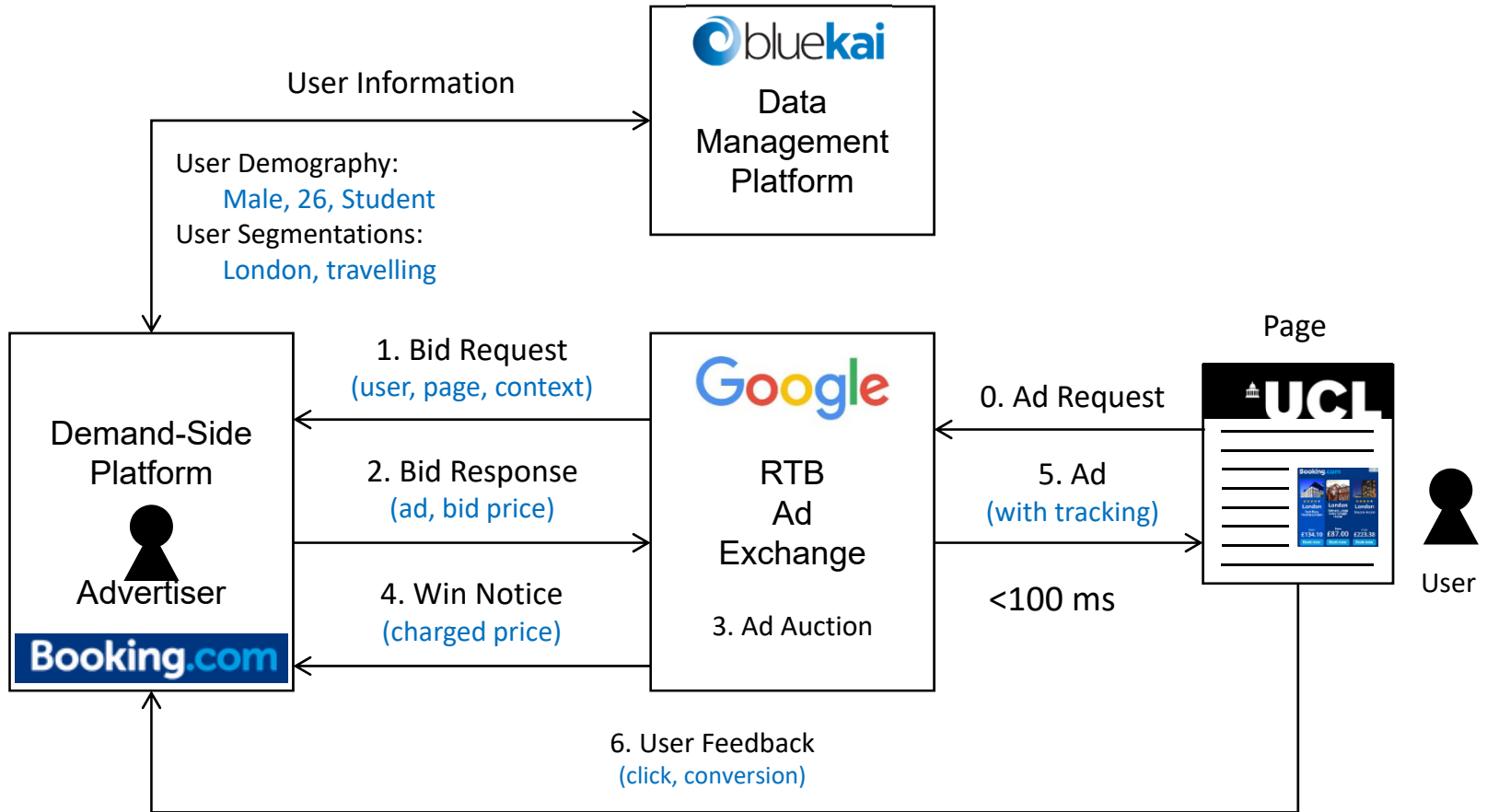


- LP: using Long term 7-day user behavior and representing the user behavior by Page-views;
- LQ: using Long term 7-day user behavior and representing the user behavior by Query terms;
- SP: using Short term 1-day user behavior and representing user behavior by Page-views;
- SQ: using Short term 1-day user behavior and representing user behavior by Query terms.

Content of This Course

- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

RTB Display Advertising Mechanism



- Buying ads via real-time bidding (RTB), 10B per day

Data of Learning to Bid

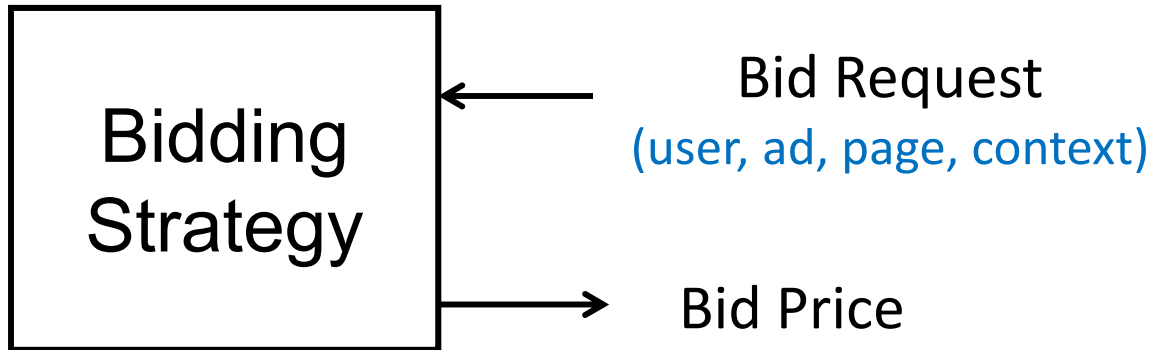
- Data

(\mathbf{x}, t)	b	w	c	y
(up, 1500×20, Shanghai, 0)	5	1	4	1
(down, 1200×25, Paris, 1)	4	1	3	0
(left, 20×1000, Los Angeles, 2)	3	0	×	×
(right, 35×600, London, 3)	0	0	×	×

- Bid request features: High dimensional sparse binary vector
- Bid: Non-negative real or integer value
- Win: Boolean
- Cost: Non-negative real or integer value
- Feedback: Binary

Problem Definition of Learning to Bid

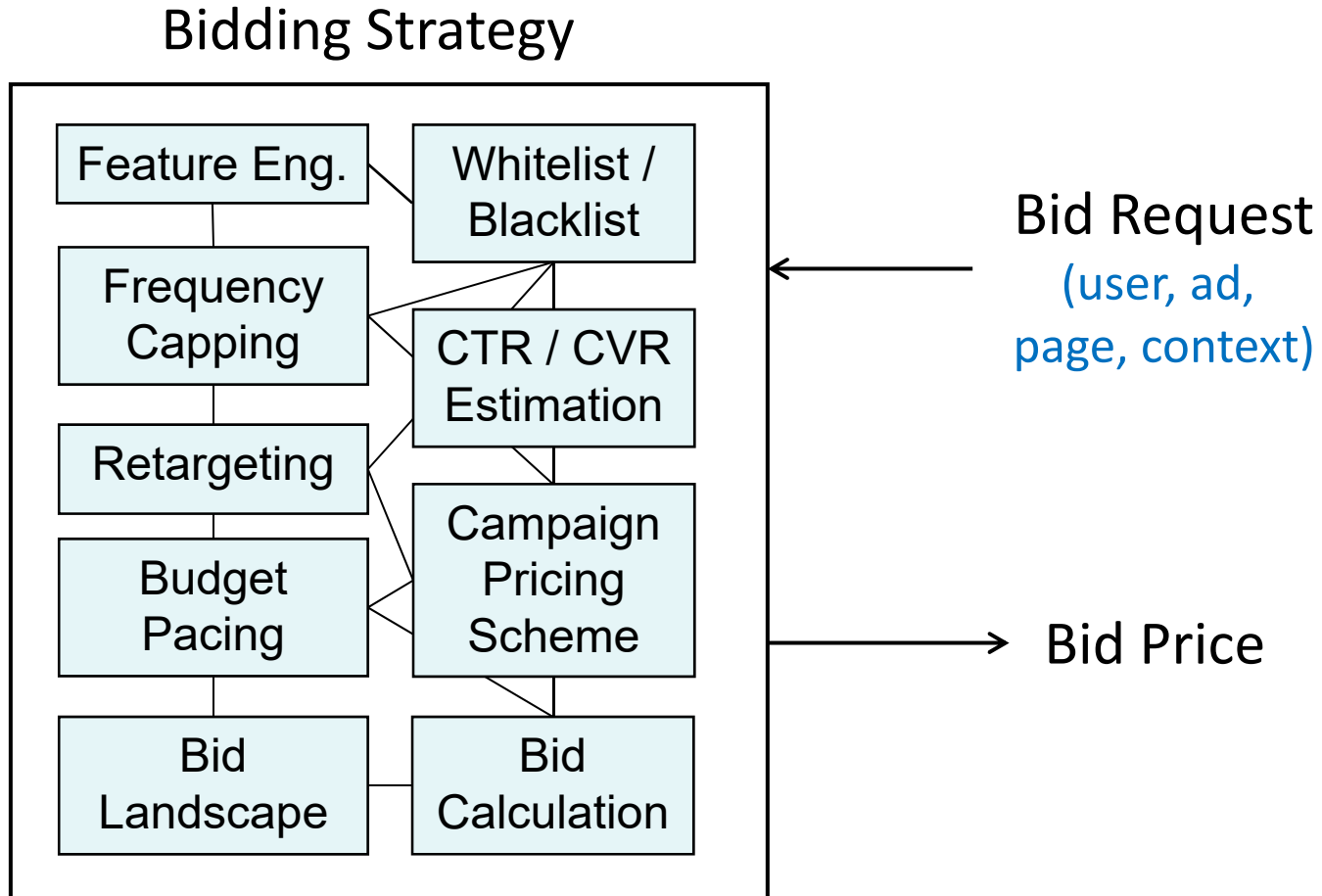
- How much to bid for each bid request?
 - Find an optimal bidding function $b(x)$



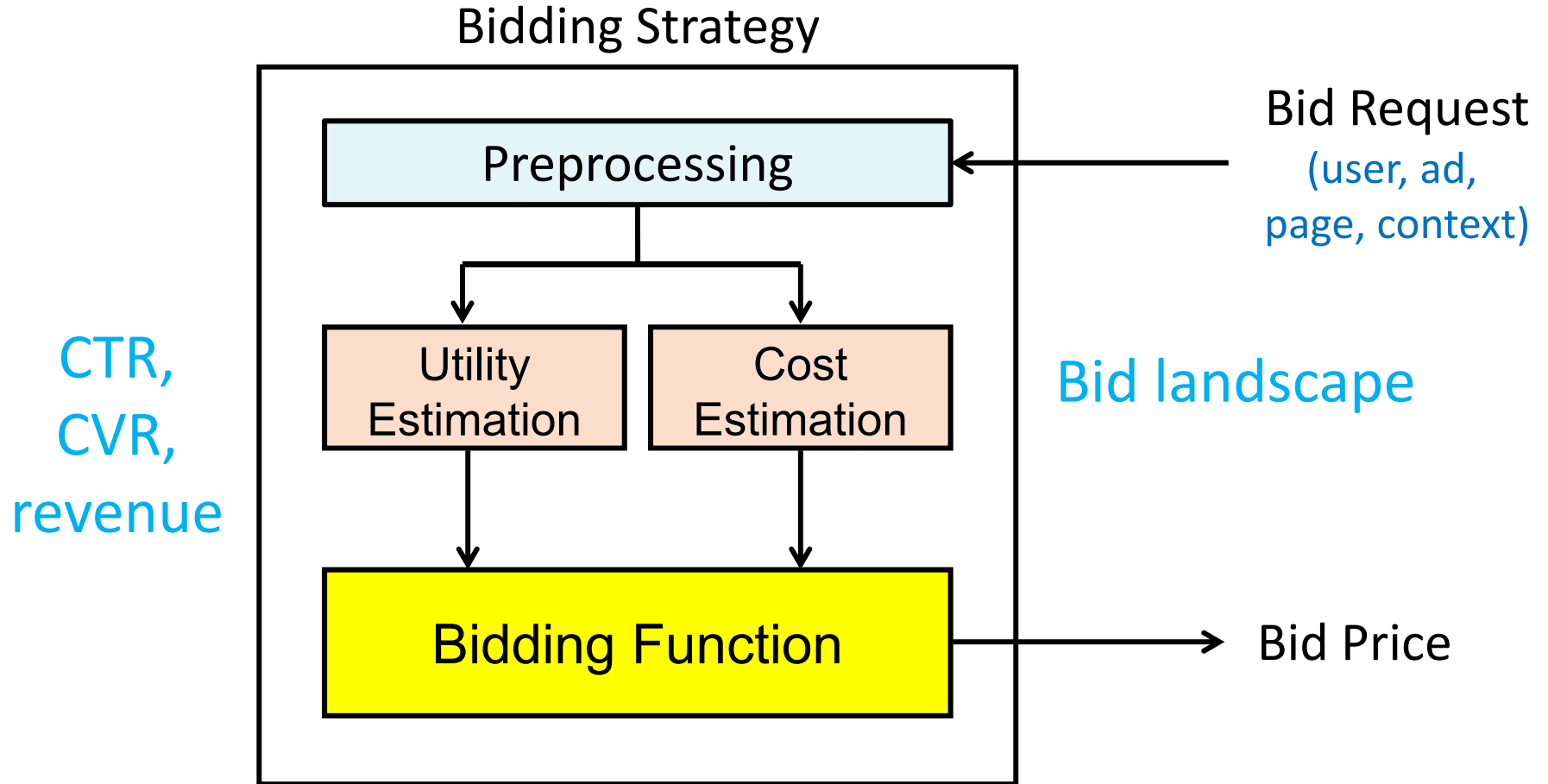
- Bid to optimize the KPI with budget constraint

$$\begin{aligned} & \max_{\text{bidding strategy}} && \text{KPI} \\ & \text{subject to} && \text{cost} \leq \text{budget} \end{aligned}$$

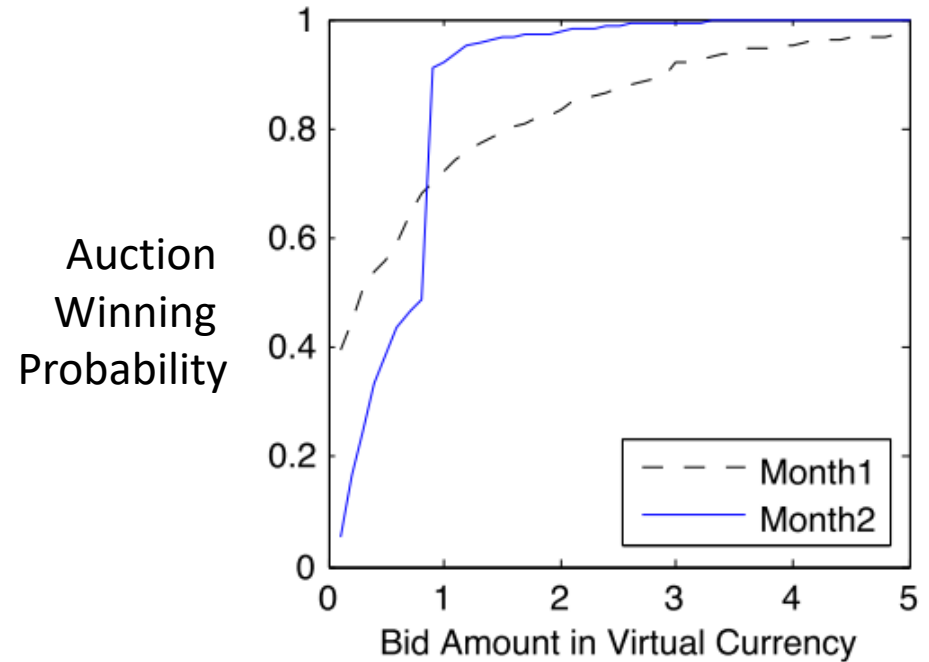
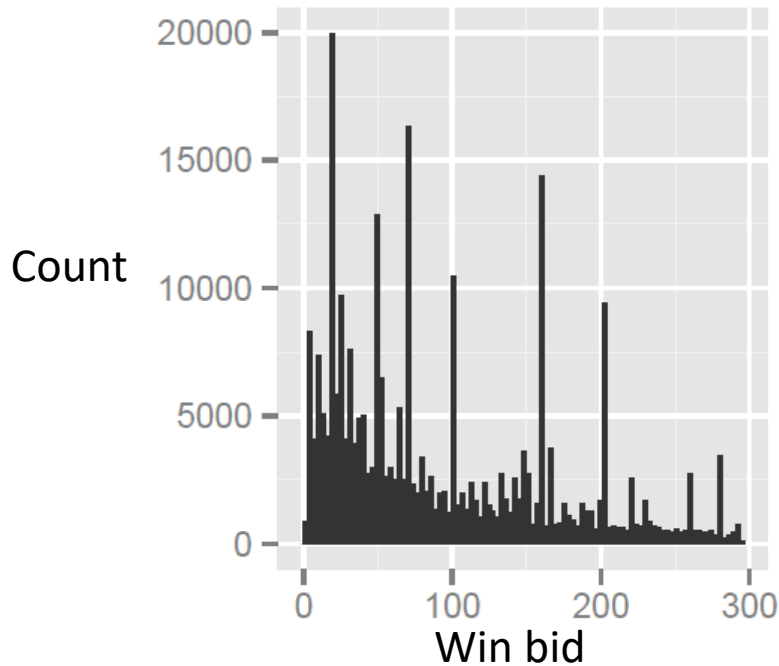
Bidding Strategy in Practice



Bidding Strategy in Practice: A Quantitative Perspective



Bid Landscape Forecasting



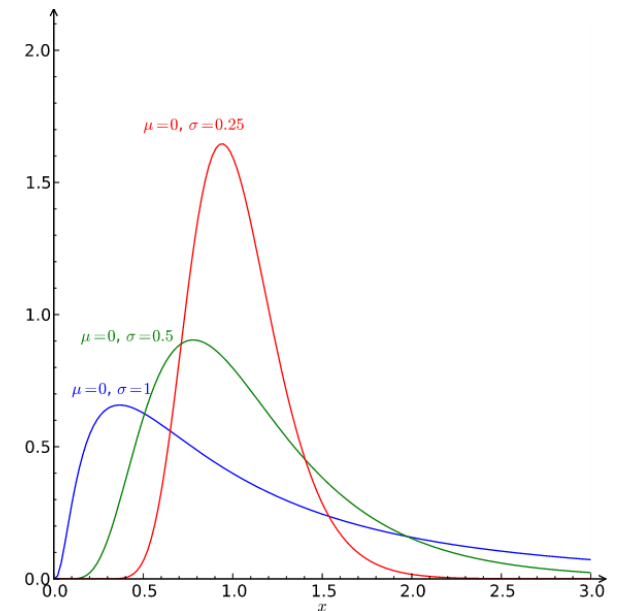
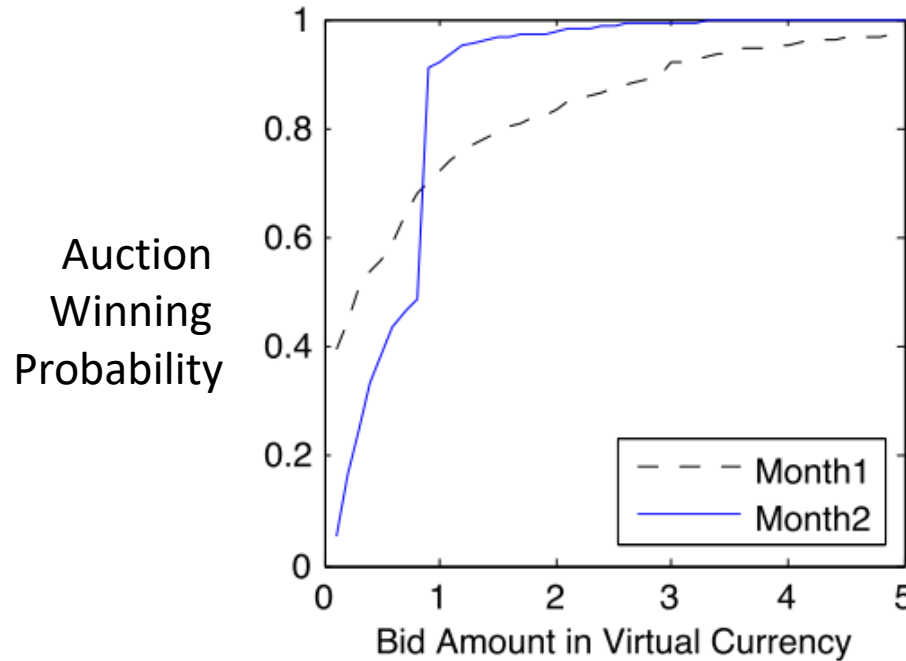
Win probability:

$$w(b) = \int_{z=0}^b p(z) dz$$

Expected cost:

$$c(b) = \frac{\int_{z=0}^b zp(z) dz}{\int_{z=0}^b p(z) dz}$$

Bid Landscape Forecasting



- Log-Normal Distribution

$$f_s(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, x > 0$$

Data Bias Problem for Bid Landscape

$$w(b) = \int_{z=0}^b p(z) dz$$

- If we directly count the probability from observed market prices

$$w_o(b_x) = \frac{\sum_{(\mathbf{x}', y, z) \in D} \delta(z < b_x)}{|D|}$$

- The estimation is unbiased since the observed market prices is always lower than the historic bid
- Counterfactual case: example of WW2 planes

Survival Model for Bid Landscape

- Kaplan-Meier Product-Limit method

$$l(b_{\mathbf{x}}) = \prod_{b_j < b_{\mathbf{x}}} \frac{n_j - d_j}{n_j} \quad w(b_{\mathbf{x}}) = 1 - \prod_{b_j < b_{\mathbf{x}}} \frac{n_j - d_j}{n_j}$$

b_i	w_i	z_i
2	win	1
3	win	2
2	lose	×
3	win	1
3	lose	×
4	lose	×
4	win	3
1	lose	×

b_j	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$w(b_j)$	$w_o(b_j)$
1	8	0	1	$1 - 1 = 0$	0
2	7	2	$\frac{5}{7}$	$1 - \frac{5}{7} = \frac{2}{7}$	$\frac{2}{4}$
3	4	1	$\frac{3}{4}$	$1 - \frac{5}{7} \frac{3}{4} = \frac{13}{28}$	$\frac{3}{4}$
4	2	1	$\frac{1}{2}$	$1 - \frac{5}{7} \frac{3}{4} \frac{1}{2} = \frac{41}{56}$	$\frac{4}{4}$

Survival Model for Bid Landscape

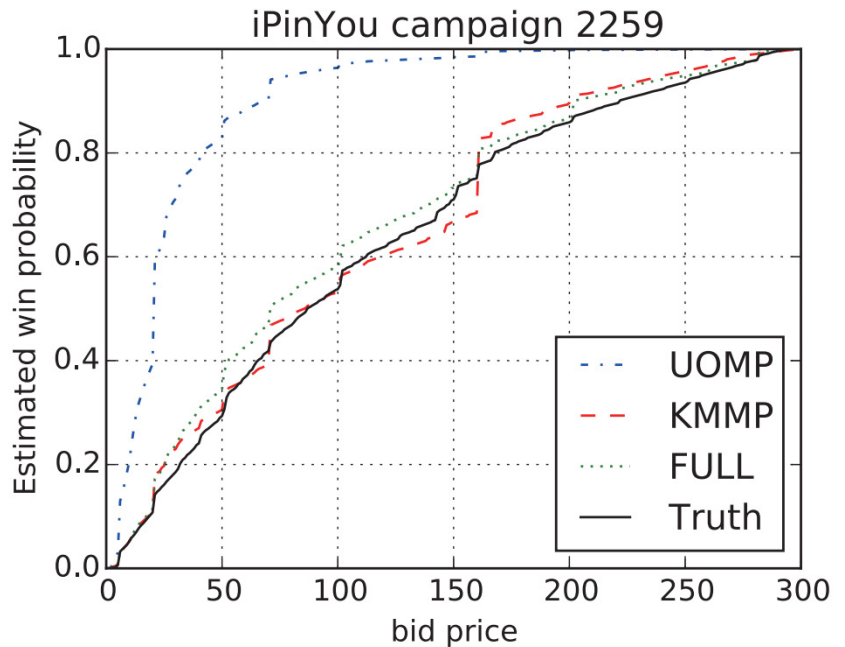
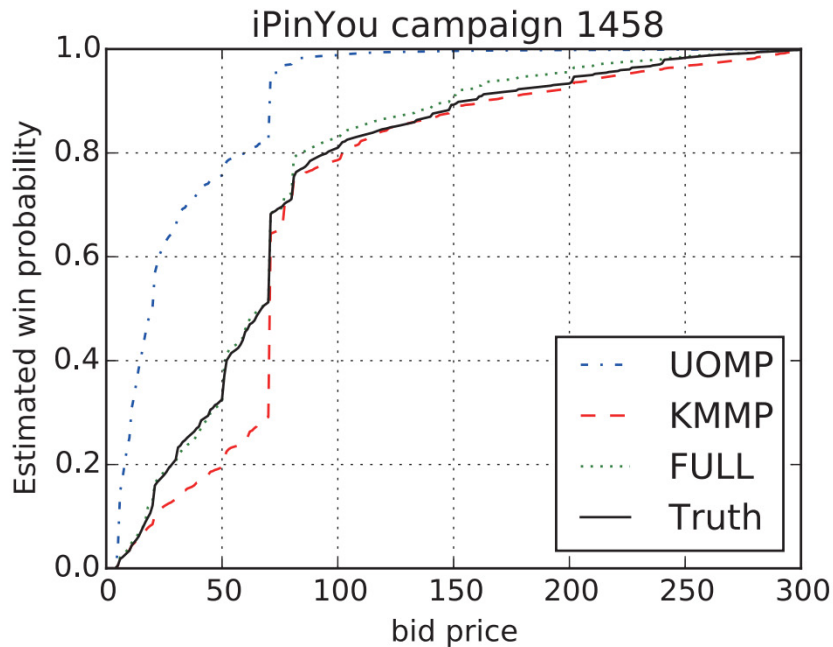
- Kaplan-Meier Product-Limit method

$$w_o(b_{\mathbf{x}}) = \frac{\sum_{(\mathbf{x}', y, z) \in D} \delta(z < b_{\mathbf{x}})}{|D|}$$

UOMP

$$w(b_{\mathbf{x}}) = 1 - \prod_{b_j < b_{\mathbf{x}}} \frac{n_j - d_j}{n_j}$$

KMMP



Bid Landscape Forecasting

- Price Prediction via Linear Regression

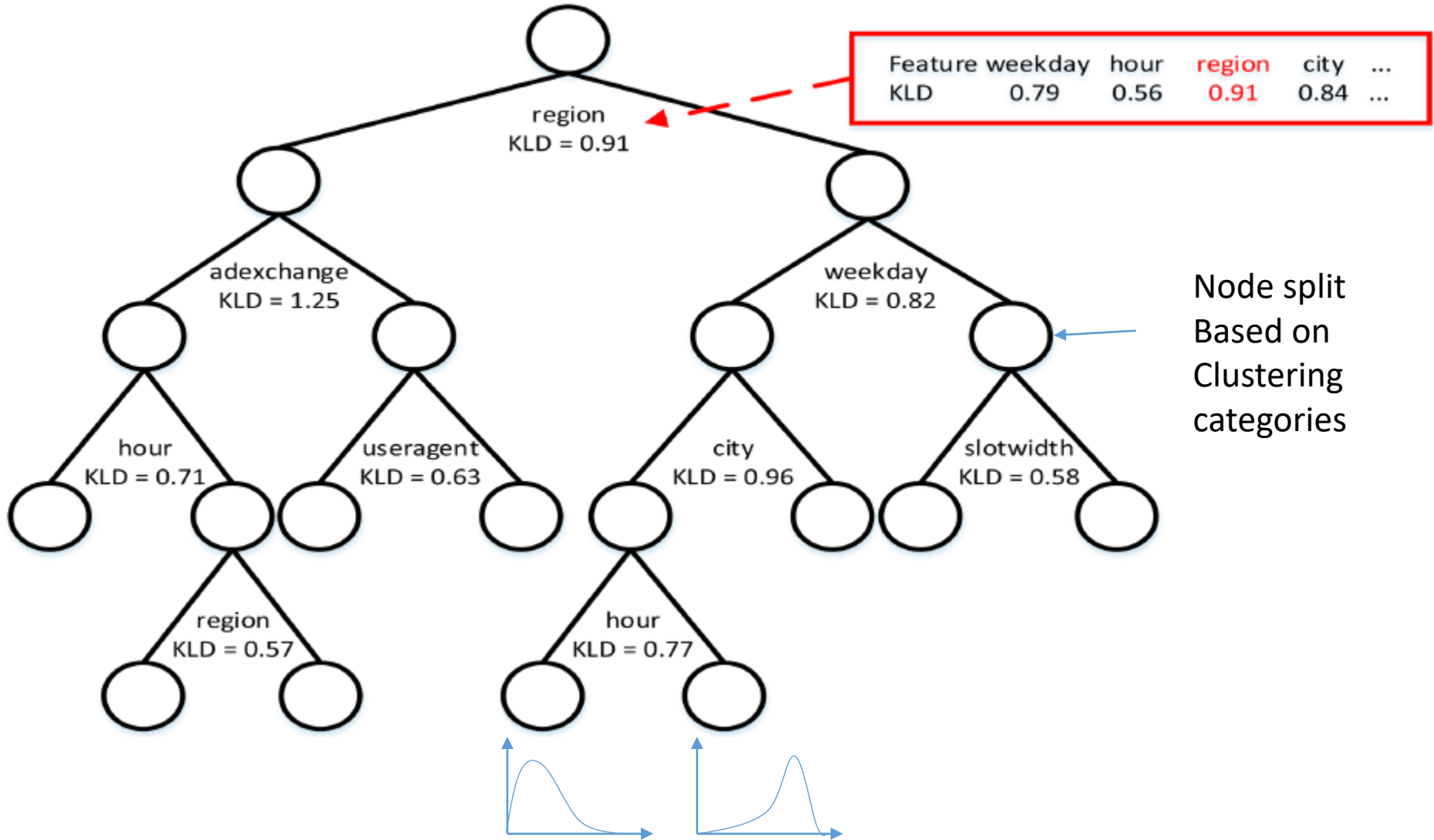
$$z = \boldsymbol{\beta}^T \mathbf{x} + \epsilon \quad \max_{\boldsymbol{\beta}} \sum_{i \in W} \log \phi \left(\frac{z_i - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma} \right)$$

- Modelling censored data in lost bid requests

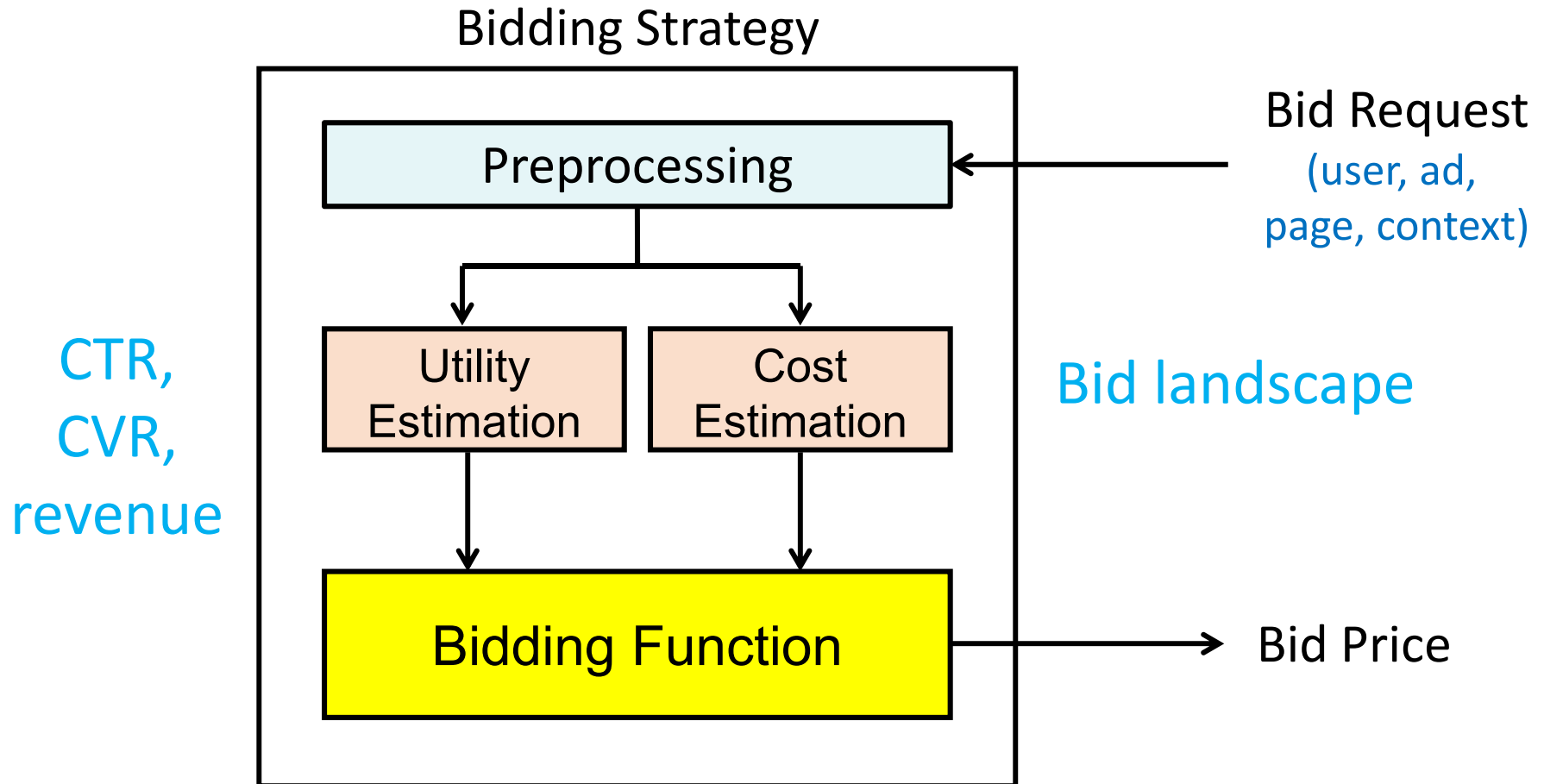
$$P(b_i < z_i) = \Phi \left(\frac{\boldsymbol{\beta}^T \mathbf{x}_i - b_i}{\sigma} \right)$$

$$\max_{\boldsymbol{\beta}} \sum_{i \in W} \log \phi \left(\frac{z_i - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma} \right) + \sum_{i \in L} \log \Phi \left(\frac{\boldsymbol{\beta}^T \mathbf{x}_i - b_i}{\sigma} \right)$$

Survival Tree Models

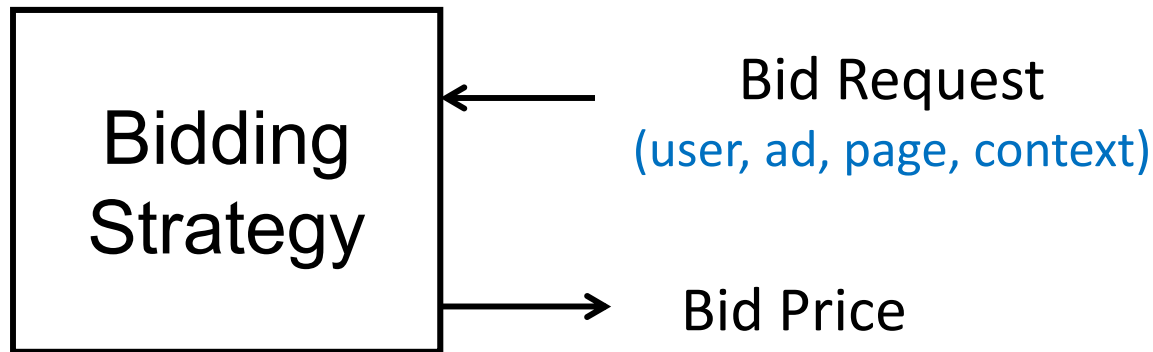


Bidding Strategy in Practice: A Quantitative Perspective



Bidding Strategies

- How much to bid for each bid request?



- Bid to optimize the KPI with budget constraint

$$\begin{array}{ll} \max & \text{KPI} \\ \text{bidding strategy} & \\ \text{subject to} & \text{cost} \leq \text{budget} \end{array}$$

Classic Second Price Auctions

- Single item, second price (i.e. pay market price)

Reward given a bid:
$$R(b) = \int_0^b (r - z)p(z)dz$$

Optimal bid:
$$b^* = \max_b R(b)$$

$$\frac{\partial R(b)}{\partial b} = (r - b)p(b)$$

$$\frac{\partial R(b)}{\partial b} = 0 \Rightarrow b^* = r \quad \text{Bid true value}$$

Truth-telling Bidding Strategies

- Truthful bidding in second-price auction
 - Bid the true value of the impression
 - Impression true value = $\begin{cases} \text{Value of click, if clicked} \\ 0, \text{ if not clicked} \end{cases}$
 - Averaged impression value = value of click * CTR
 - Truth-telling bidding:

$$\text{bid} = r_{\text{conv}} \times \text{CVR} \quad \text{or} \quad \text{bid} = r_{\text{click}} \times \text{CTR}$$

Truth-telling Bidding Strategies

$$\text{bid} = r_{\text{conv}} \times \text{CVR} \quad \text{or} \quad \text{bid} = r_{\text{click}} \times \text{CTR}$$

- Pros
 - Theoretic soundness
 - Easy implementation (very widely used)
- Cons
 - Not considering the constraints of
 - Campaign lifetime auction volume
 - Campaign budget
 - Case 1: \$1000 budget, 1 auction
 - Case 2: \$1 budget, 1000 auctions

Non-truthful Linear Bidding

- Non-truthful linear bidding

$$\text{bid} = \text{base_bid} \times \frac{\text{predicted_CTR}}{\text{base_CTR}}$$

- Tune base_bid parameter to maximize KPI
- Bid landscape, campaign volume and budget indirectly considered

$$\begin{array}{ll} \max & \text{KPI} \\ \text{bidding strategy} & \\ \text{subject to} & \text{cost} \leq \text{budget} \end{array}$$

ORTB Bidding Strategies

- Direct functional optimisation

$$b()_{\text{ORTB}} = \arg \max_{b()} N_T \int_{\theta} \theta w(b(\theta)) p_{\theta}(\theta) d\theta$$

winning function CTR

$$\text{subject to } N_T \int_{\theta} b(\theta) w(b(\theta)) p_{\theta}(\theta) d\theta \leq B \leftarrow \text{budget}$$

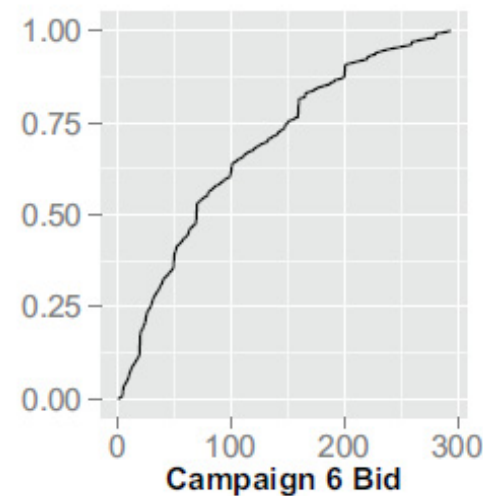
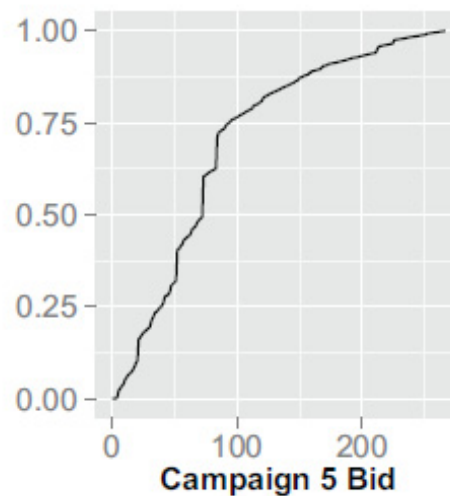
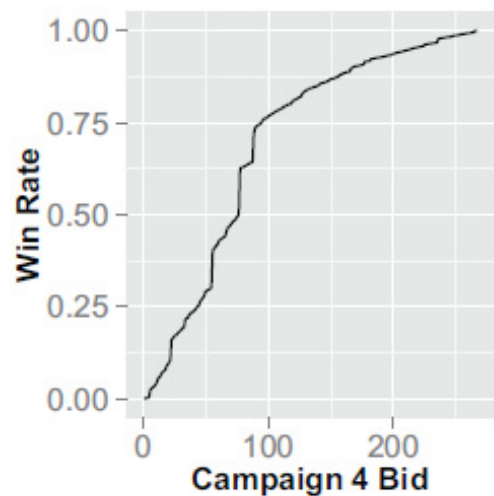
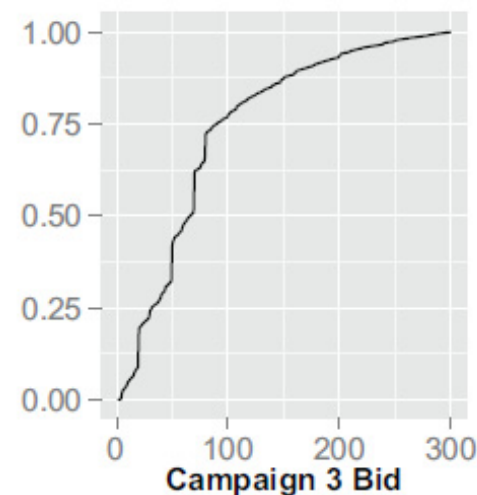
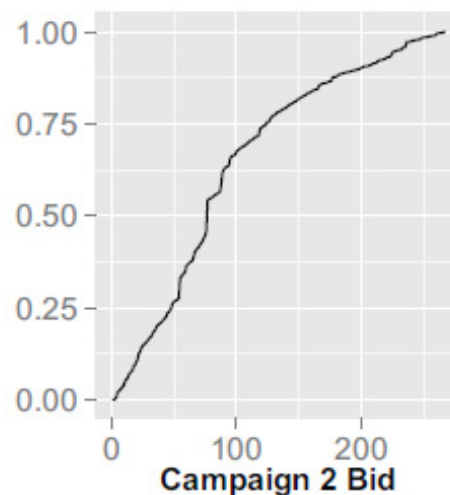
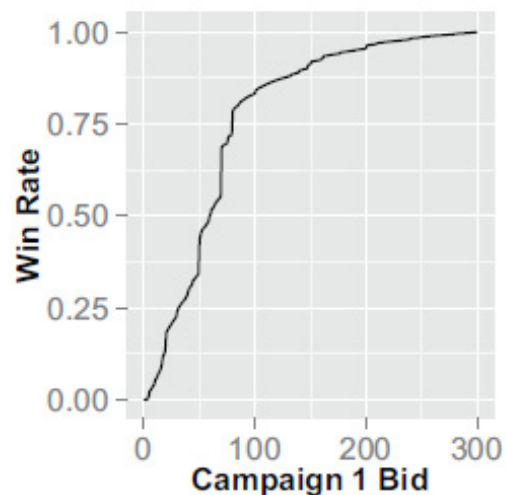
Est. volume bidding function cost upperbound

- Solution: Calculus of variations

$$\mathcal{L}(b(\theta), \lambda) = \int_{\theta} \theta w(b(\theta)) p_{\theta}(\theta) d\theta - \lambda \int_{\theta} b(\theta) w(b(\theta)) p_{\theta}(\theta) d\theta + \frac{\lambda B}{N_T}$$

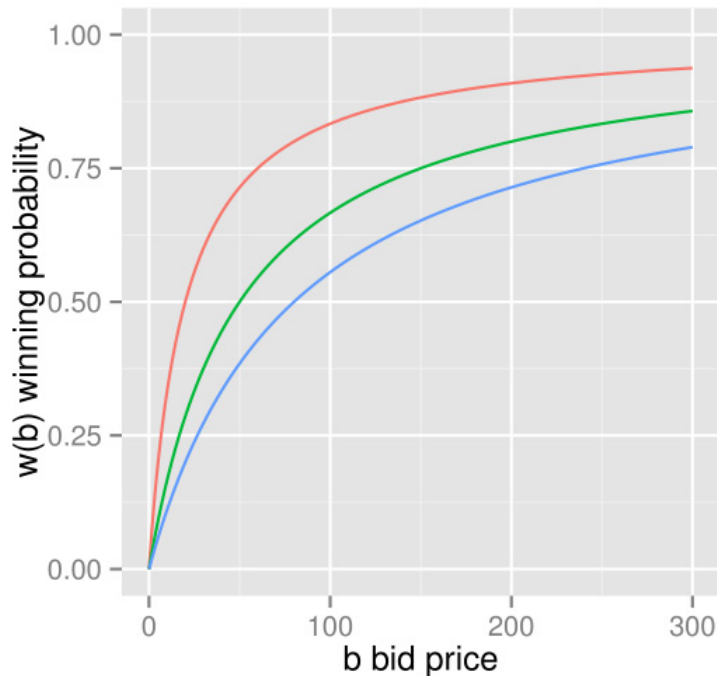
$$\frac{\partial \mathcal{L}(b(\theta), \lambda)}{\partial b(\theta)} = 0 \quad \Rightarrow \quad \lambda w(b(\theta)) = \left[\theta - \lambda b(\theta) \right] \frac{\partial w(b(\theta))}{\partial b(\theta)}$$

Bid Landscape: $w(\text{bid})$

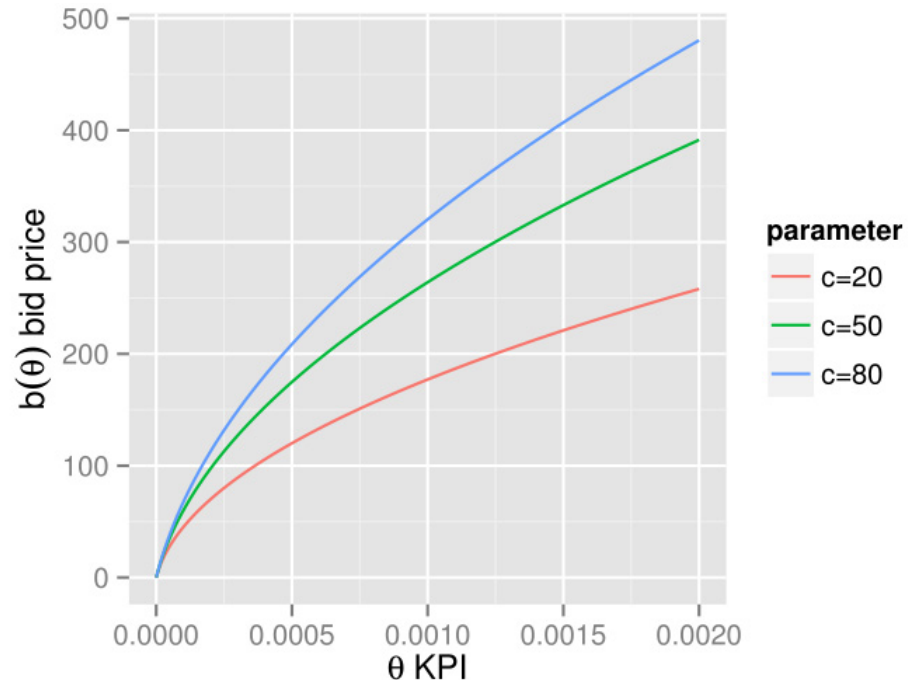


Optimal Bidding Strategy Solution

$$\lambda w(b(\theta)) = \left[\theta - \lambda b(\theta) \right] \frac{\partial w(b(\theta))}{\partial b(\theta)}$$



(a) Winning function 1.

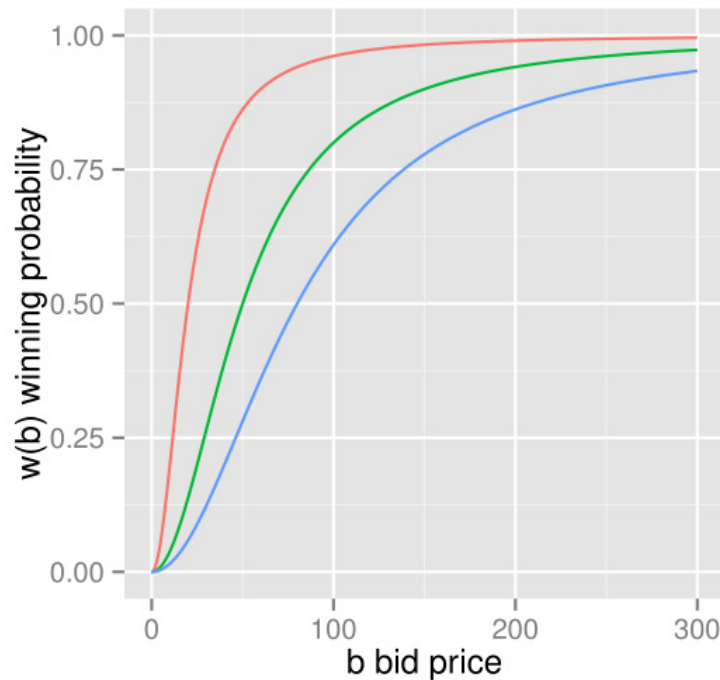


(b) Bidding function 1.

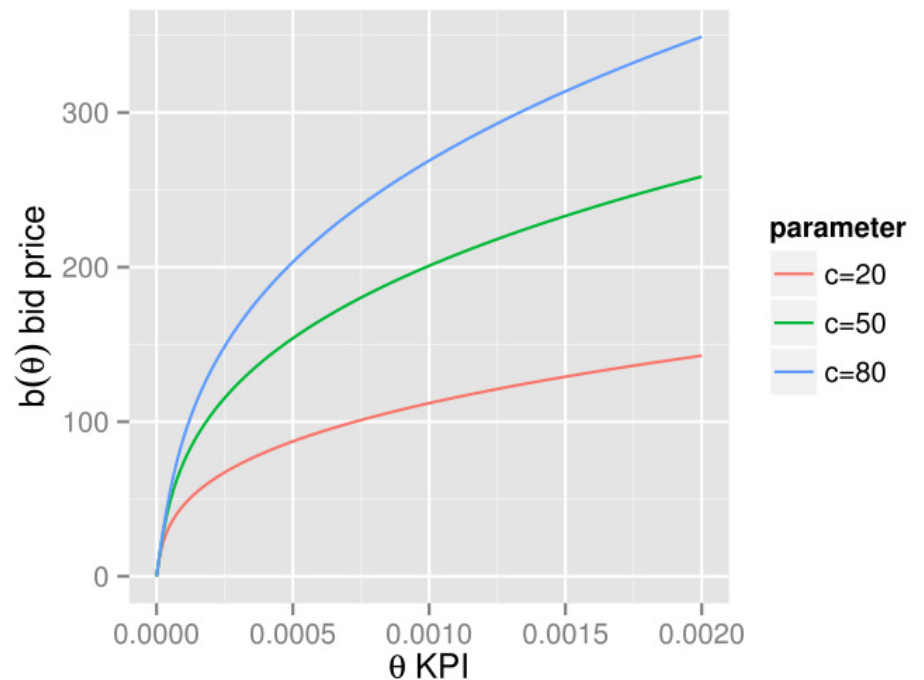
$$w(b(\theta)) = \frac{b(\theta)}{c + b(\theta)} \quad \Rightarrow \quad b_{\text{ORTB1}}(\theta) = \sqrt{\frac{c}{\lambda} \theta + c^2} - c$$

Optimal Bidding Strategy Solution

$$\lambda w(b(\theta)) = \left[\theta - \lambda b(\theta) \right] \frac{\partial w(b(\theta))}{\partial b(\theta)}$$



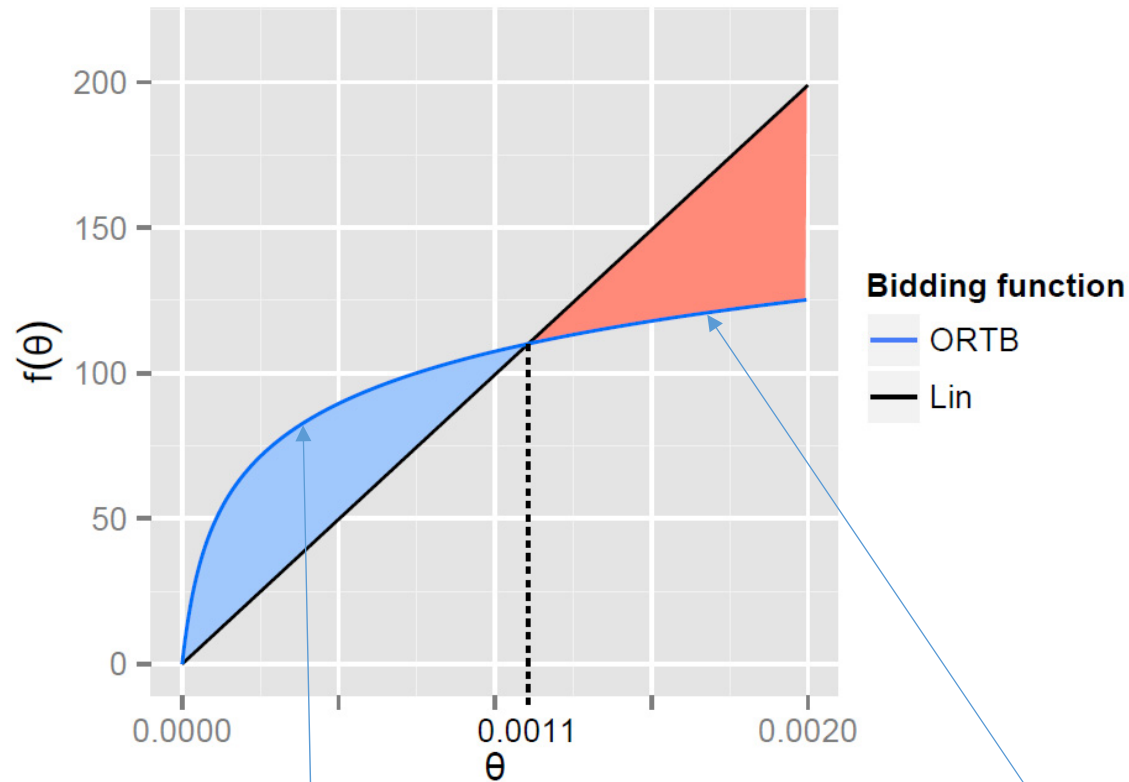
(a) Winning function 2.



(b) Bidding function 2.

$$w(b(\theta)) = \frac{b^2(\theta)}{c^2 + b^2(\theta)} \rightarrow b_{\text{ORTB2}}(\theta) = c \cdot \left[\left(\frac{\theta + \sqrt{c^2 \lambda^2 + \theta^2}}{c \lambda} \right)^{\frac{1}{3}} - \left(\frac{c \lambda}{\theta + \sqrt{c^2 \lambda^2 + \theta^2}} \right)^{\frac{1}{3}} \right]$$

Optimal Bidding Strategy: the Analysis



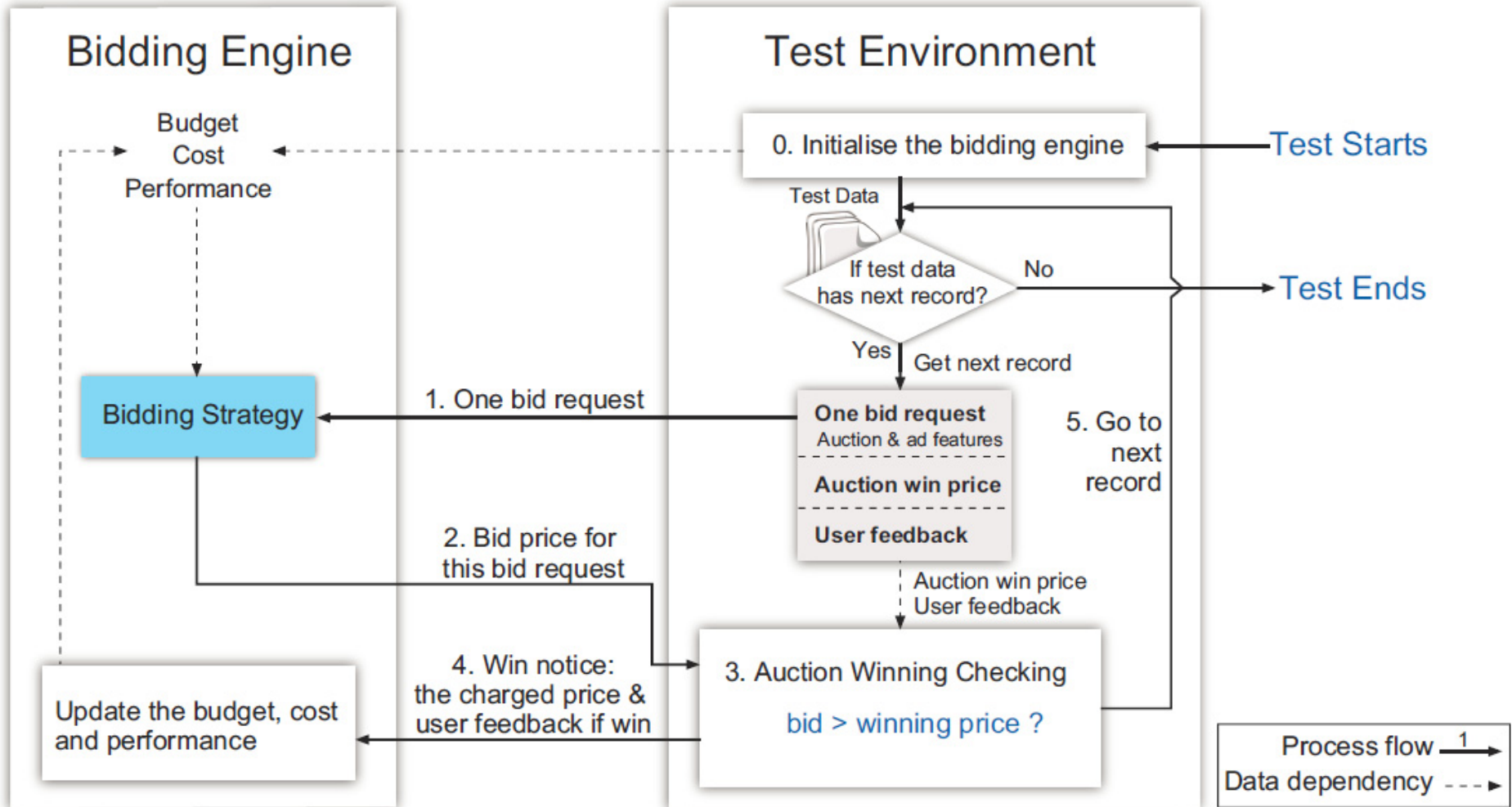
Slight increase at low bids is more effective

Thus reduce the bids at high CTR or CVR

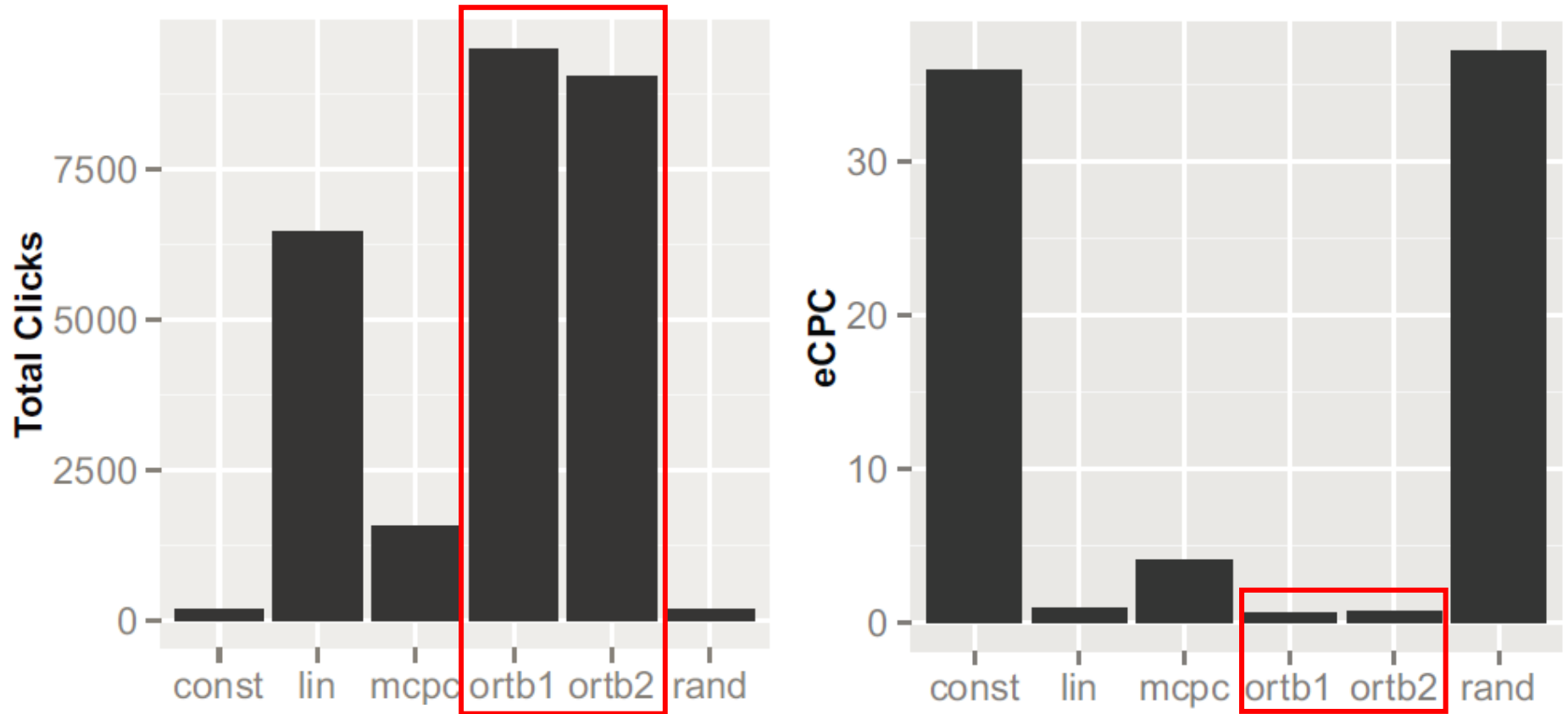
Experiment

- We used iPinYou's dataset
 - 1-<http://data.computational-advertising.org>
 - 9 Campaigns, 15M impressions, 11K clicks, 935 conversions
- Evaluated bidding strategies
 - **Const**: Constant
 - **Rand**: Random
 - **Mcpc**: Bidding based on advertiser's given max eCPC [Chen et al. 2011]
 - **Lin**: Linear to pCTR [Perlich et al. 2012]
 - **ORTB1, ORTB2**: Optimal bidding strategies with two forms of winning rate functions

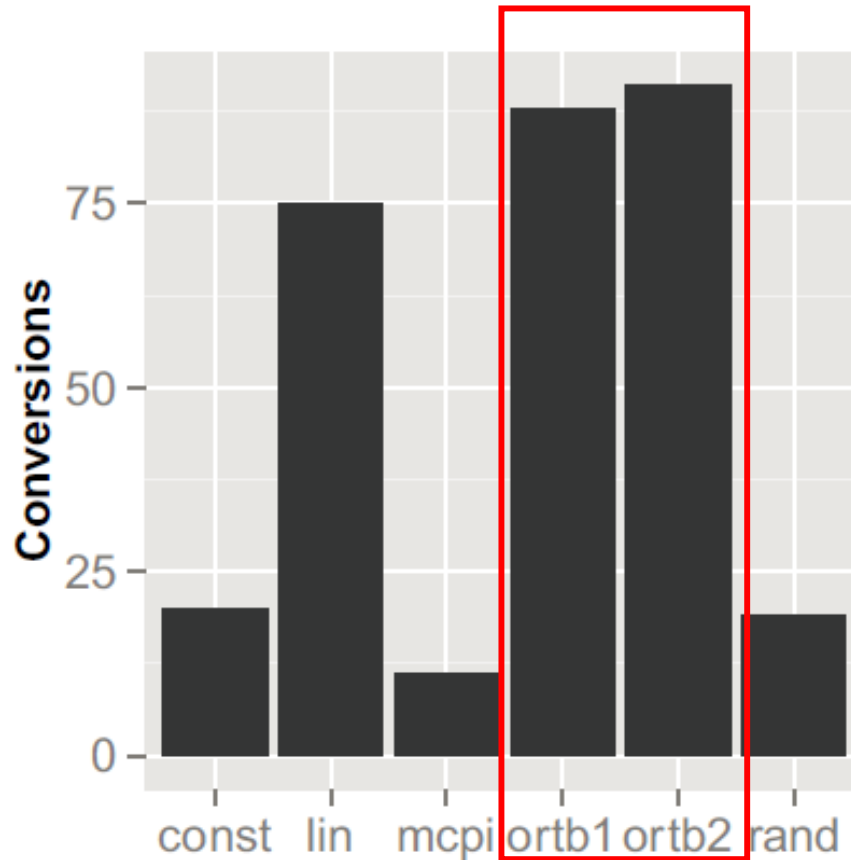
Offline Test Evaluation Flow



Overall performance: Optimizing Clicks



Overall performance – Optimizing Conversions



Unbiased Optimization

- Bid optimization on ‘true’ distribution

$$\arg \max_{b(\cdot)} T \int_{\mathbf{x}} f(\mathbf{x}) w(b(f(\mathbf{x}))) p_x(\mathbf{x}) d\mathbf{x}$$

$$\text{subject to } T \int_{\mathbf{x}} b(f(\mathbf{x})) w(b(f(\mathbf{x}))) p_x(\mathbf{x}) d\mathbf{x} = B$$

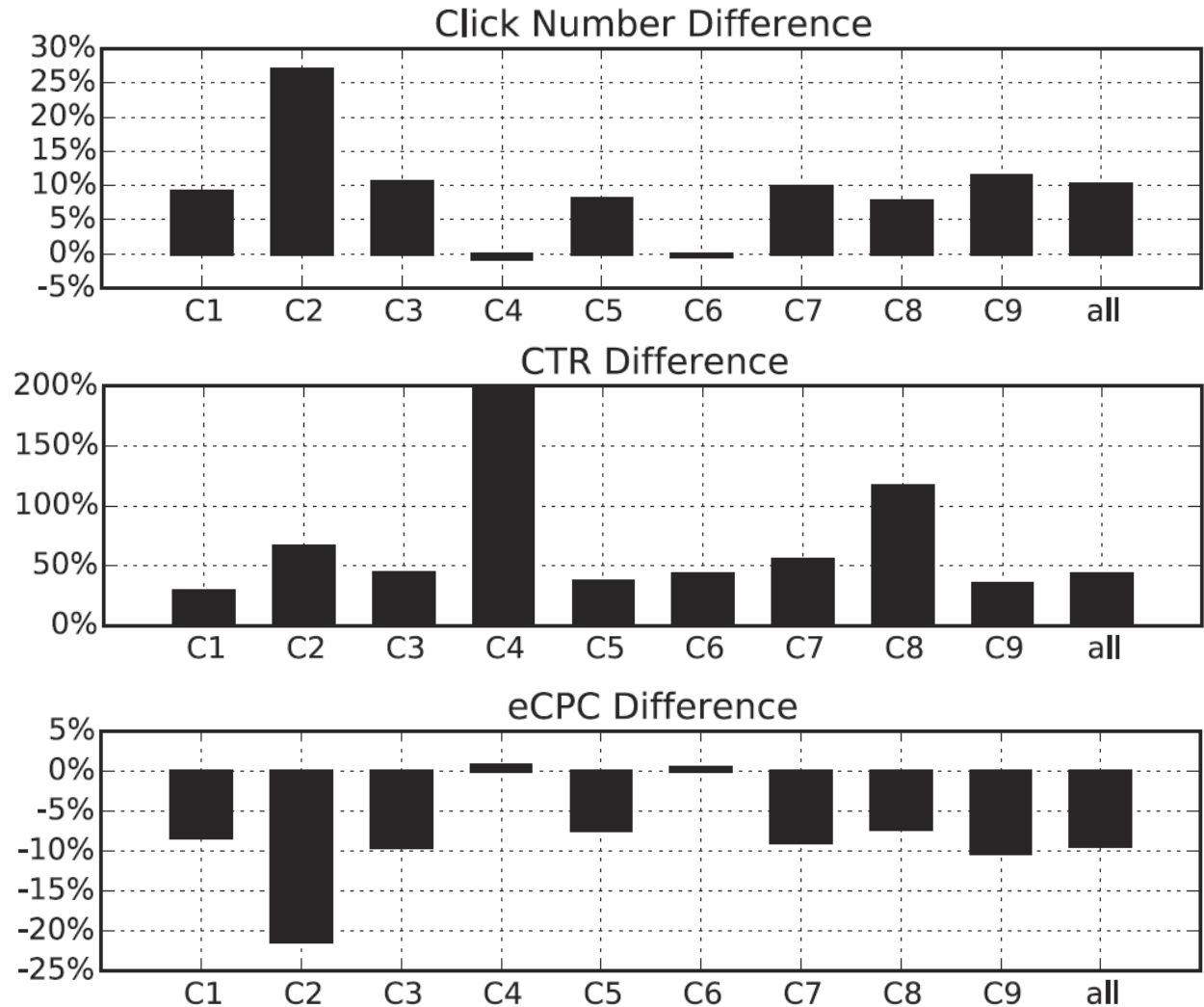
- Unbiased bid optimization on biased distribution

$$\arg \max_{b(\cdot)} T \int_{\mathbf{x}} f(\mathbf{x}) w(b(f(\mathbf{x}))) \frac{q_x(\mathbf{x})}{w(b_x)} d\mathbf{x}$$

$$\text{subject to } T \int_{\mathbf{x}} b(f(\mathbf{x})) w(b(f(\mathbf{x}))) \frac{q_x(\mathbf{x})}{w(b_x)} d\mathbf{x} = B$$

Unbiased Bid Optimization

A/B Testing
on Yahoo!
DSP.



Content of This Course

- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

Fraud

- Reported by Interactive Advertising Bureau's (IAB) in 2015
- Ad fraud is costing the U.S. marketing and media industry an estimated \$8.2 billion each year
- \$4.6 billion, or 56%, of the cost to “invalid traffic”, of which 70% is performance based, e.g., CPC and CPA, and 30% is CPM based.

An Display Ad Example

How do you know the user is a human or a robot?

大陆



河南省公安厅彻查“封丘36人入警 35人身份不合规”

中封丘县公安局的36名受训人员，35人是公安局内部的文职或临时人员，与“民警必须具备公务员身份”的国家规定不符，引发该局内部

- 上海至成都沿江高铁提上日程 串联长江沿线22城市
- 2016号歼-20原型机曝光 已滑行测试(图)
- 日媒：中国或派万吨海警船巡钓鱼岛 打消耗战
- 外媒：中国开始研制隐身武装直升机 预计2020年交付
- 习近平关于中美关系的十个判断
- 住建部黑臭水沟整治工作指南：9成百姓满意才能达标
- 陕西：职校“校长”让女学生陪酒 学校被撤除
- 揭秘“团团伙伙”的武钢漩涡和落马高官

国际



巴塞罗那200万人游行 呼吁加泰罗尼亚独立(图)

- 李炜光：收税是不公平的恶？
- 许章润：超级大国没有纯粹内政
- 刘昉献：国外政党联系群众的路径研究

时局观



民革中央副主席：中共从未否定国民党抗战作用

- 施芝鸿：文革基础上搞改革致一个时期市场官场乱象
- 朱维群回应争议：尊重民族差异而不强化
- 伊协副会长：穆斯林不应因宗教功修忽视社会责任

领袖圈



奥巴马54岁啦，当7年总统人苍老了头发也白了



海绵城市 未来之城
水危机：青岛告急
探访中国绿化博览会
帝都吸引华人首富
凤凰房产 诚邀加盟

谈华山论剑与中国精神
黑龙江创新驱动三步棋
《印记》之江城夜未眠
办公环境搜查令
圈层生活尽在凤凰网

精彩视频

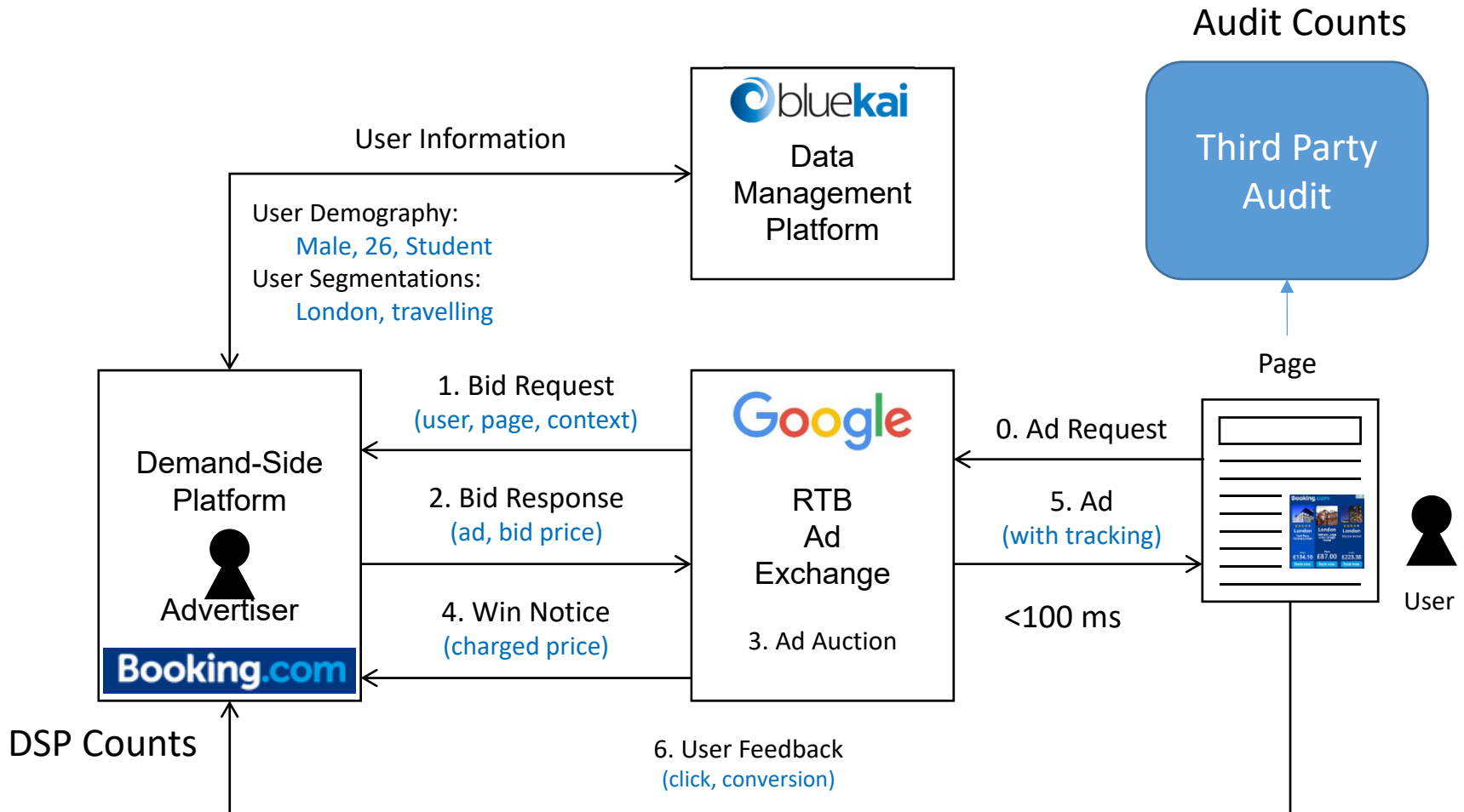
凤凰联播台



菲媒曝菲律宾军演针对中国 直指南海生命线

播放数：2602282

Leverage Third Party to Audit



- Typically, the counts of the DSP and Audit should be close
 - Say $\pm 5\%$

A Good Story of Fraud Fighters

- <http://www.rtbchina.com/inside-google-s-secret-war-ad-fraud.html>



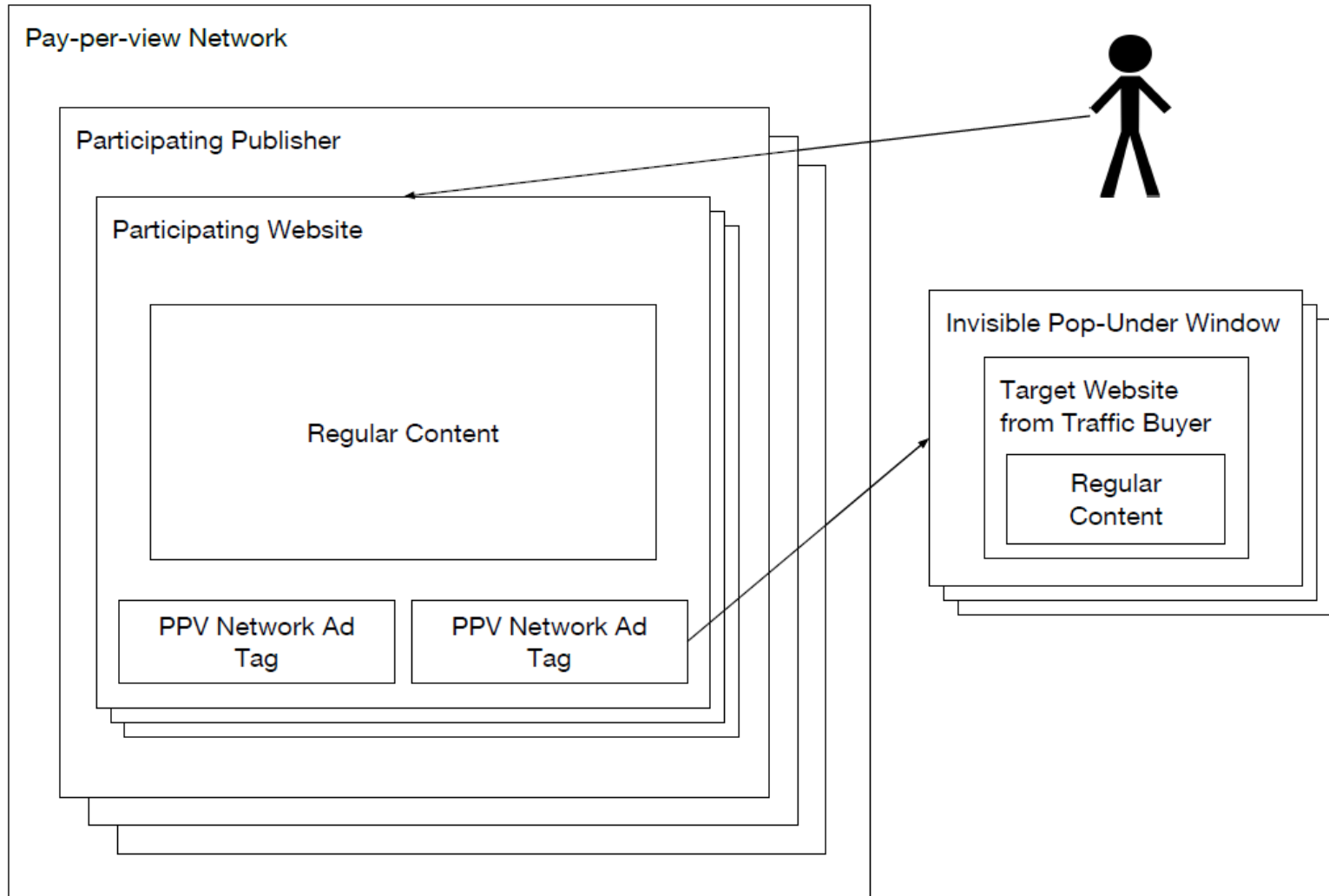
Ad Fraud Types

- Impression fraud
 - where the fraudster generates fake bid requests, sells them in ad exchanges, and gets paid when advertisers buy them to get impressions
- Click fraud
 - where the fraudster generates fake clicks after loading an ad
- Conversion fraud
 - where the fraudster completes some actions, e.g., filling out a form, downloading and installing an app, after loading an ad

Ad Fraud Sources

- Publisher driven: pay-per-view network
- User/robot driven: botnet

Pay-Per-View (PPV) Networks



Possible Methods to Avoid PPV for Advertisers

- Viewport size check: valid impressions will not be displayed in a 0x0 viewport, which is invisible to users
- A referrer blacklist, which checks if the traffic is from the PPV networks
- A publisher blacklist, which avoids buying traffic from publishers who participate in the PPV networks

Botnets

- Botnets are usually built with compromised end users' computers.
 - These computers are installed with one or multiple software packages, which run autonomously and automatically.
 - Adware

Adware Examples

The screenshot shows a web browser window with the address bar displaying <http://www.pcrisk.com/>. The page features a navigation menu with links for HOME, REMOVAL GUIDES, NEWS, BLOG, FORUM, TOP ANTI-SPYWARE, and TOP ANTI-MALWARE. A prominent advertisement for "Call for Great Tech Support" is overlaid on the right side, featuring a toll-free number 1-855-565-3218 and a "TOLL FREE" badge. Below this, a "Download" button is visible. A "Important Message" box in the center-left contains a question mark icon and text stating: "Your download manager might be outdated. Click here to download the upgrade." Below this message, the text "Ads by CheckMeUp" is visible. Two orange arrows point from the "Important Message" box and the "Call for Great Tech Support" ad towards the "New Removal Guides" section. This section is divided into two columns. The left column lists "Online Video Promoter Adware" with a description: "Furthermore, Online Video Promoter tracks Internet browsing activity and collects various information. IP addresses, websites visited, search queries, pages viewed, and other collected data might contain personally identifiable details, thus, having Online Video Promoter installed on your system may consequently result in serious privacy issues or even identity theft. It is worth mentioning that other adware applications distributed using the bundling method (e.g., UnknownFile, 1Player, CorAdviser, GetitHD, HQ Video Pro) are very similar to Online Video Promoter. Every adware promises user to enable various useful functions, however, neither of them are actually useful - their true purpose is to generate e...". The right column lists "iShopper Ads" (described as "What is more, iShopper collects diverse softw..."), "YTDownloader Adware" (described as "On top of that, as any other potentially unwa..."), "Threat Finder Ransomware" (described as "The 'help decrypt' files"), "UnknownFile Adware", and "1Player Adware". Each item in the right column includes a small thumbnail image and a label "Adware".

A Few Ways to Detecting Botnets

- Signature based detection, which extracts software / network package signature from known botnet activities
- Anomaly detection of traffic
- DNS based detection, which focuses on analyzing DNS traffic which is generated by communication of bots and the controller
- **Mining based detection**, which uses Machine Learning techniques to cluster or classify botnet traffic

Data Mining based Fraud Detection

- Ad fraud detection is usually an unsupervised learning problem and it is difficult to capture the ground-truth
- Fully unsupervised learning
 - Detect the fraud based on the revealed web structures and human heuristics
- Semi-supervised learning
 - Detect the fraud by training a predictor based on a very small labeled data and large unlabeled data

Ad Fraud Detection with Co-visit Networks

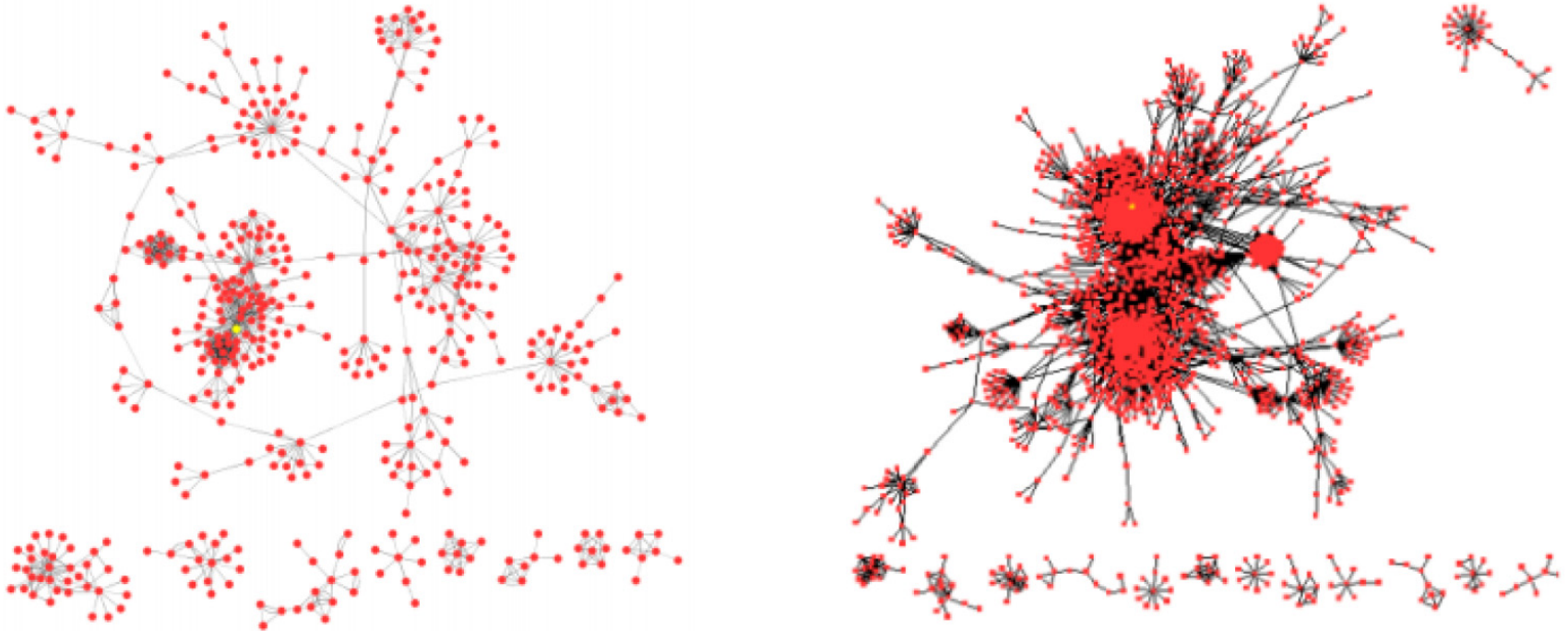
- Define a bipartite graph between users (browsers) and websites

$$G = \langle B, W, E \rangle$$

- B : users
 - W : websites
 - E : the edge indicating whether the user has visit the website over a specified time period
-
- The co-visit network is based on G

$$G_W^n = \langle V_W \subseteq W, E = (x, y) : x, y \in W, [\Gamma_G(x) \cap \Gamma_G(y)] / \Gamma_G(x) \geq n \rangle$$

Co-Visit Network Examples



- The co-visit networks of Dec 2010 (left) and Dec 2011 (right) reported by Stitelman et al. [2013].

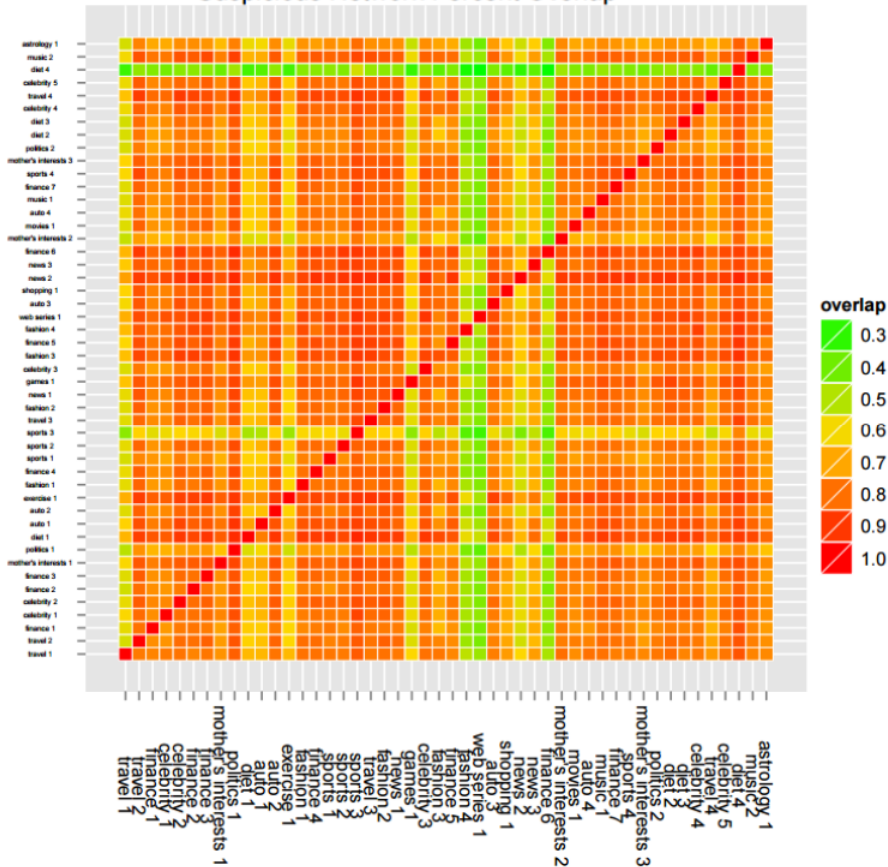
Co-Visit Network for Fraud Detection

- Intuition: two websites' user overlap is normally very small
 - High dimensional random vectors are almost vertical (i.e. with cosine close to 0)

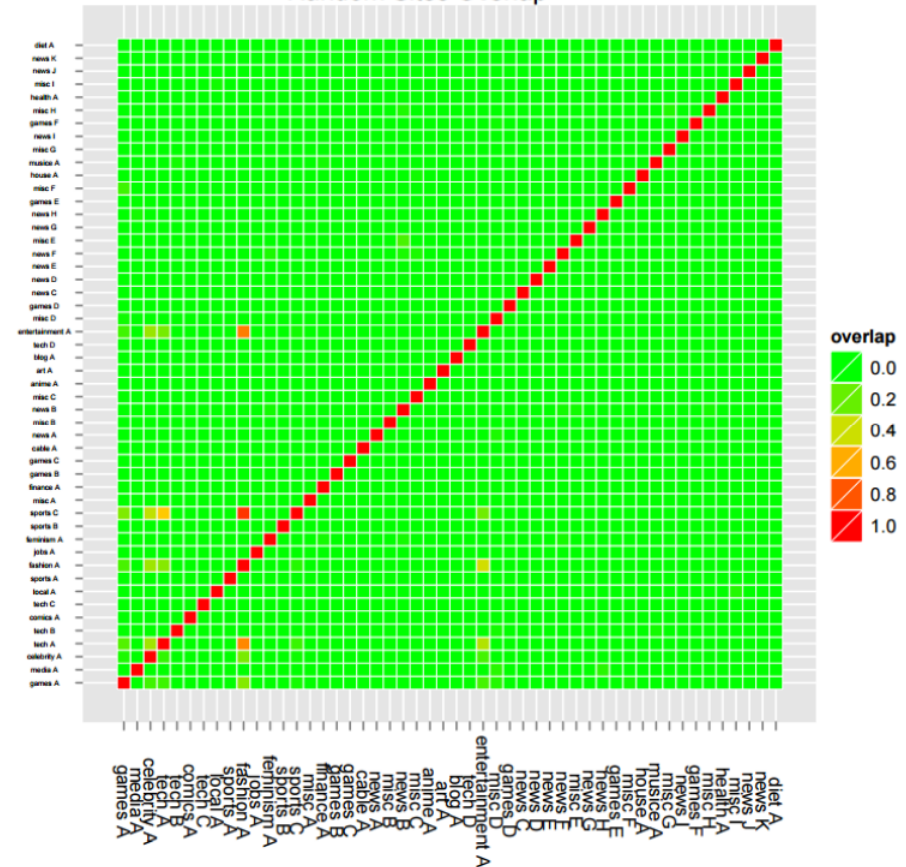
Co-Visit Network for Fraud Detection

- Intuition: two websites' user overlap is normally very small

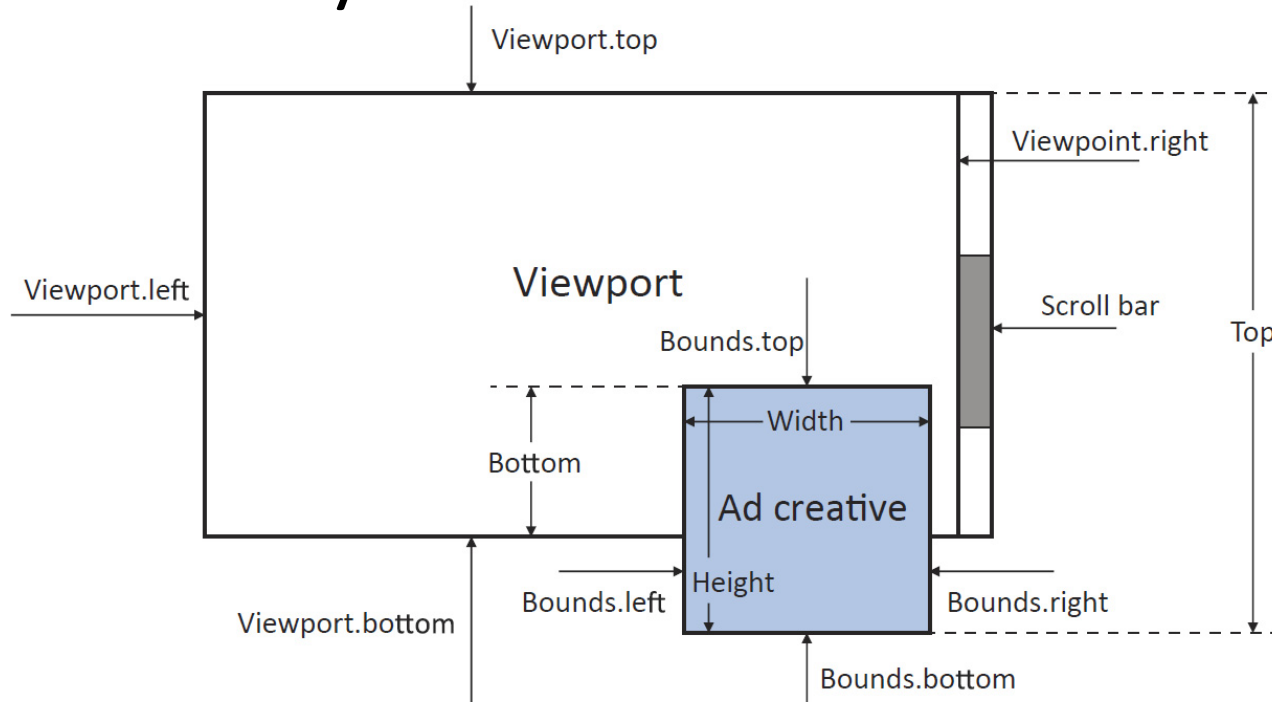
Suspicious Network Percent Overlap



Random Sites Overlap



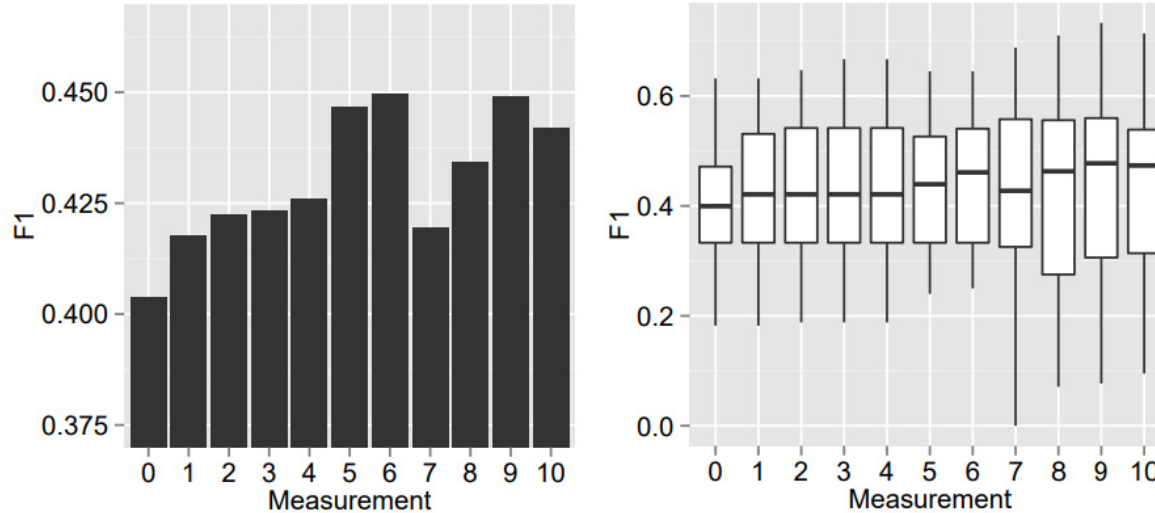
Viewability Methods



We developed a javascript to tracking each user's behavior on browsing a displayed ad

- Pixel percentage tracking: The displayed pixel percentage for rectangle ad creative in the viewport
- Exposure time tracking: The exposure time is associated with a pixel percentage threshold.

Viewability Methods



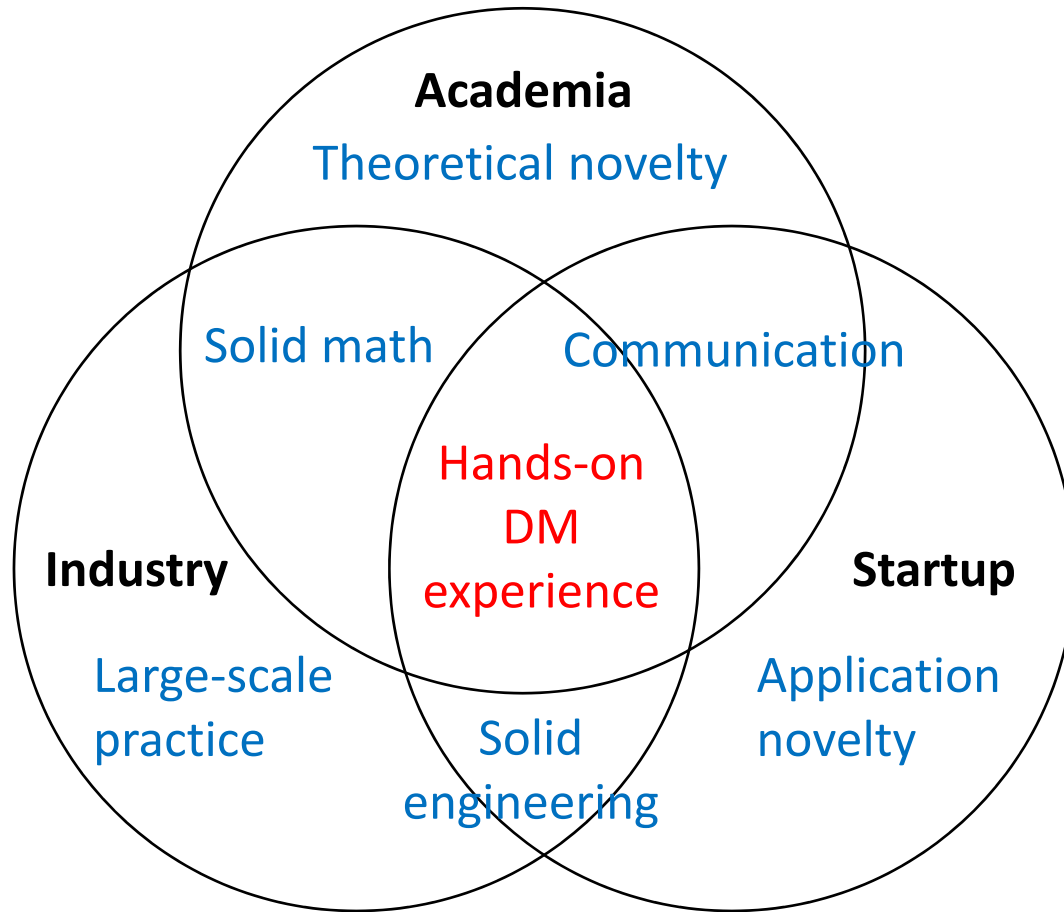
meas.	pixel	time	meas.	pixel	time	meas.	pixel	time
0	0%	0	4	100%	(0,2)	8	50%	[4,+)
1	0%	(0,2)	5	50%	[2,4)	9	75%	[4,+)
2	50%	(0,2)	6	75%	[2,4)	10	100%	[4,+)
3	75%	(0,2)	7	100%	[2,4)			

- Results: (pixel \geq 75%, time \geq 2s) provided the highest average F1 score and median F1 score

Summary of EE448

1. Data Mining Intro
2. Fundamentals of Data
3. Basic DM Algorithms
4. Supervised Learning 1
5. Supervised Learning 2
6. Supervised Learning 3
7. Supervised Learning 4
8. Unsupervised Learning
9. Search Engines
10. Ranking Information Items
11. Recommender Systems
12. Computational Ads
13. Behavioral Targeting
14. Poster Session

We focus on hands-on DM



- Get familiar with various data mining applications.
- Play with the data and get your hands dirty!

Thank You!

Weinan Zhang, Ph.D.
Assistant Professor



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

John Hopcroft Center for Computer Science
Dept. of Computer Science & Engineering
Shanghai Jiao Tong University